The Dissertation Committee for Vinh Huy Pham Certifies that this is the approved version of the following dissertation:

Computer Modeling of the Instructionally Insensitive Nature of the Texas Assessment of Knowledge and Skills (TAKS) Exam

Committee:

Walter M. Stroup, Supervisor

Guadalupe Carmona-Dominguez

James P. Barufaldi

Daniel I. Bolnick

Emin T. Ulug

Computer Modeling of the Instructionally Insensitive Nature of the Texas Assessment of Knowledge and Skills (TAKS) Exam

by

Vinh Huy Pham, Bachelor's of Science in Biology and Chemsitry

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

August, 2009

Dedication

This dissertation is dedicated to all the teachers in the world with the hope that they will find the work contained within this dissertation pertinent to their practice.

Acknowledgements

No one can hope to stand on their own and be able to achieve anything of value relative to the larger society. This dissertation was made possible through the support of all my friends and family how helped to make me the person I am today. I would also like to thank my advisors, Dr. Walter Stroup and Dr. Lupita Carmona, for giving me the freedom to explore and discover what was important to me when it comes to my research. I would like to thank my graduate program coordinators, Amy Always and Jennifer Wagner, for keeping me from going off course. Lastly, I would like to thank my fiancé, Kendall Reynolds, for loving me with all my peculiar quirks and habits.

Computer Modeling of the Instructionally Insensitive Nature of the Texas Assessment of Knowledge and Skills (TAKS) Exam

Publication No._____

Vinh Huy Pham, Ph.D.

The University of Texas at Austin, 2009

Supervisor: Walter M. Stroup

Stakeholders of the educational system assume that standardized tests are transparently about the subject content being tested and therefore can be used as a metric to measure achievement in outcome-based educational reform. Both analysis of longitudinal data for the Texas Assessment of Knowledge and Skills (TAKS) exam and agent based computer modeling of its underlying theoretical testing framework have yielded results that indicate the exam only rank orders students on a persistent but uncharacterized latent trait across domains tested as well as across years. Such persistent rank ordering of students is indicative of an instructionally insensitive exam. This is problematic in the current atmosphere of high stakes testing which holds teachers, administrators, and school systems accountable for student achievement.

Table of Contents

List of Figures

List of Tables

List of Equations

CHAPTER 1: Introduction

Population biology, in its broadest senses, is about understanding the interaction between aspects of the individual and aspects of the population as they play out in relation to an environment. In switching the focus of my graduate work from formal biology to science education, I initially thought my focus was to develop a way to better teach biology, as what is sometimes called "content". Only later did I begin to see the issues surrounding the application of psychometrics to education, especially as related to "high stakes" testing, as a special instance of the relating of aspects of the individual to aspects of the population as situated in the environment of schooling. Credible models in population biology make clear both their assumptions and the ways in which such models would be expected to fit with actual, or plausible, real world data. Given the status high stakes testing has in education as well as some perplexing testing results that we were getting from research projects focused on supporting innovations in mathematics education (Stroup et al., 2007), I began to develop a sense that our understanding of current psychometric practices, especially as applied to high stakes tests like the Texas Assessment of Knowledge and Skills (TAKS) might be advanced by a similar clarification of the assumptions and the fit between the models being used and the testing data.

At a top most level, this dissertation develops out of an attempt to engage psychometrics in ways similar to the kinds of engagements with modeling

assumptions and fit with data that I had used in my graduate work in biology. Especially as we worked to explain some of the patterns and anomalies in high stakes test results observed across a number of projects, I began to wonder how well the actual behavior of the tests could be accounted for from a careful examination of assumptions and by the use of various modeling techniques. Accordingly, the following research questions are going to be addressed by this dissertation:

- Is it possible to build credible models of high stakes tests that are highly attentive to the assumptions informing the use of two principle approaches to test construction: Classical Testing Theory and Item Response Theory, and that fit well with, or simulate effectively, the behavior of high stakes tests in Texas?
- What implications or issues related to current psychometric practice, especially as associated with the use of IRT for high stakes test construction and analyses, are made visible from the modeling approaches pursued in this dissertation?

John Dewey was one of the first educational thinkers to compare education to biology. In his seminal work, **Democracy and Education**, Dewey writes:

> It is the very nature of life to strive to continue in being. Since this continuance can be secured only by constant renewals, life is a self-renewing process. What nutrition and reproduction are to physiological life, education is to social life. This education consists primarily in transmission through communication. Communication is

> a process of sharing experience till it becomes a common possession. (p.11).

Education can be thought of as the process by which immature and less able members of society become more mature and able through interacting with each other and experts. In addition to being about individuals, education serves a social function at the population level. “As societies become more complex in structure and resources, the need for formal or intentional teaching and learning increases.” (Dewey, 1916, p. 11). Systems of education were created to regulate and frame both what and how individuals learn by providing a “special social environment which shall especially look after nurturing the capacities of the immature.” (Dewey, 1916, p. 27). Individuals who are the recipients of education are called students to indicate their status as immature and less able and “are not regarded as social members in full and regular standing.” (Dewey, 1916, p. 63). When students have completed their education, it is hoped that they will become productive individuals in their society and continue to perpetuate their society as a culture. How the process of education is framed psychometrically and the effectiveness of the models currently used in this effort are the foci of this work.

Chapter 2 begins the process of analysis and modeling of psychometry by making clear the assumptions of both the more traditional Classical Testing Theory and contrasting these with the assumptions of Item Response Theory (IRT). While CTT is no longer popular, with most high stakes standardized tests using IRT as their foundation, it is discussed in the dissertation for the

comparative value it brings to the discussion. The discussion will also bear on the Texas Assessment of Knowledge and Skills (TAKS) exam and its specific underlying framework, the one parameter logistic (1PL) model of IRT as a prototypical standardized test. **Chapter 3** proceeds with a discussion of the data sources used in the dissertation, including how they were processed in order to commence to analysis. Also discussed in the chapter are the types of analyses used to discern the mechanistic principles that govern test score results and ultimately the underlying psychometric model. These analyses can be classified into three general categories: student behavior, item behavior, and the interaction between student and test. The remainder of the dissertation uses a two pronged approach to examine the psychometric model of the TAKS exam. The first approach in **Chapter 4** deals with the analysis of real world longitudinal data to see the actual results of the TAKS exam implementation and how they might relate to the assumptions of the IRT-1PL framework and its psychometric model. The second approach in **Chapter 5** uses agent based computer modeling to examine the assumptions of the IRT-1PL framework and attempts to explain how the results in the previous chapter can be situated within the theoretical context. **Chapter 6** cross validates both approaches by using real world data for students in the model through confirmatory analyses. **Chapter 7** wraps up the dissertation with a discussion of the conclusions drawn from all the analyses as well as the possible implications and consequences for the current educational environment. It ends with possible future trajectories that could be pursued to elaborate beyond the story told in this dissertation.

CHAPTER 2: Theoretical Testing Frameworks and the TAKS Exam

During any renewal process, there must be a guarantee that the replication process will maintain a high level of fidelity to the original source. In the field of education, this guarantee is ensured by various assessments that can be used to judge the quality of education as well as allowing for the comparison of students against each other and the established standards of quality (Hashway, 1998). The attaining of the standards of quality in education is termed achievement. Unfortunately, achievement is not a physical object like the length of a room that can be accurately measured without debate, but rather is a mental latent trait that must be approximated via tests (Kline, 2005). The approximation of mental latent traits by a test represents the first two major assumptions of testing: that there exists such a mental latent trait called achievement and that its magnitude can be made known as an observable score (Hashway, 1998). Society generally desires a measure of achievement that is universal in nature. Such measurements would allow for comparisons to be made both within the current population of students being assessed as well as across years with different populations of students. The desire for a universal measurement is the reason why psychometricians have created standardized tests.

Standardized tests allow for the uniform measurement of student achievement relative to the standards of learning and have been in use since the early 1900's (Brooks, 1922). There has been a strong focus in the United States

on standardized tests and improving test scores within the educational community ever since the No Child Left Behind Act (NCLB) was passed into federal law in 2001. NCLB mandates that all public schools must annually administer a state-wide standardized test to their students. Teachers and school systems are then held accountable for the results on the standardized test based on the premise of Adequate Yearly Progress (AYP) which requires that each year schools perform better than their previous year. Schools that fail to meet AYP are sanctioned and continued sanctioning could result in the closing and restructuring of the school. For this reason, standardized testing has become synonymous with “high stakes testing”. NCLB is situated in the standards-based education reform movement which holds that setting high standards and establishing measurable goals can improve educational outcomes (Ellis, 2003).

NCLB grew out of the concern that U.S. students were underachieving when compared to international standards of achievement (U.S. Department of Education, 2008). Various indicators of international achievement such as the Trends in International Mathematics and Science Study (TIMSS) have shown that U.S. students were falling behind other developed nations (Gonzales et al., 2008). NCLB attempts to both remedy this supposed deficiency in our students as well as hold teachers and schools systems accountable for their perceived failure. With a federal mandate to implement the mass standardized testing of our students whose scores could impact the welfare of teachers and school systems, it is important to ascertain if and how well these high stakes standardized tests can measure student achievement. To do this would require

an understanding of the mechanistic principles of how standardized tests work. This chapter will examine the two major modern theoretical testing frameworks used in the construction of standardized tests, Classical Testing Theory (CTT) and Item Response Theory (IRT), and then specifically target the Texas Assessment of Knowledge and Skills (TAKS) exam as a prototypical standardized test.

Classical Testing Theory

The first major modern theoretical testing framework to be developed was Classical Testing Theory (CTT). CTT assumes that there is a mental latent trait that can be measured as a true score with some level of error associated with it as represented by **Equation 2.1** (Allen & Yen, 1979).

$$\underset{\text{(Observed Score)}}{X} = \underset{\text{(True Score)}}{T} + \underset{\text{(Error)}}{E}$$

Equation 2.1 CTT's linear model of scores

Theoretically, the true score is the mean score that an examinee would get on a test if the examinee took that same test an infinite number of times. This is obviously not feasible with any testing population since taking a test even once is more than enough for most students. Multiple administrations of the same test to a student will also cause learning effects to appear in the test scores over time, and cause a change in the student's true score value (Kline, 2005).

Certain assumptions were made to get around this dilemma. One of these assumptions is that that the error of measurement (E) must be unsystematic and therefore random and uncorrelated to the true score (T) and should be normally distributed about T. Therefore:

$$\sigma_{TE} = 0$$

Equation 2.2 CTT's assumption that the covariance of T and E must be zero

Mathematically, it can be proven that the true and error scores developed from multiple administrations of a single test to one examinee can also hold true over a single administration to multiple examinees (Allen & Yen, 1979). In this case,

the same variance for the error of measurement (σ_E^2) that would have been derived from one examinee taking a test an infinite number of times can now be generalized from an entire population taking the same test only once by **Equation 2.3**.

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2$$

Equation 2.3 Classical Testing Theory's linear model of variance

The shift from an individual to the population allowed CTT to be used in a more practical large scale manner. The proof of how this shift is possible is as follows:

A. $\sigma_X^2 = \sigma_{T+E}^2$	by **Equation 2.1**
B. $\sigma_{T+E}^2 = \sigma_T^2 + 2\sigma_{TE} + \sigma_E^2$	Algebraically $(x+y)^2 = x^2 + 2xy + y^2$
C. $\sigma_X^2 = \sigma_T^2 + 2\sigma_{TE} + \sigma_E^2$	Combining B into A
D. $\sigma_X^2 = \sigma_T^2 + \sigma_E^2$	Based on **Equation 2.2**

Equation 2.3 makes no sense when applied to an individual since an examinee's T should not vary, but when applied to a population, σ_T^2 represents the variance of the true scores for that population while σ_E^2 represents the variance of the error of measurement for that same population.

There are many limitations to using the CTT framework. One of the assumptions of CTT is that the standard error of measurement does not differ between examinees. Regardless of whether an examinee is a low, middle or high achieving student, the standard error of measurement will remain the same. Secondly, the longer a test is the more reliable it becomes and is a matter of sampling. The more items there are on a test, the more the statistics generated

from it will hold true across the infinite population of possible items. Third, the sample taking the test must be representative of the population if the conclusions drawn are to be confidently generalized back to the population. Fourth, true scores must be measured on an interval scale and be normally distributed. Otherwise “test developers must convert scores, combine scales and do a variety of things to the data to ensure that these assumptions are met” (Kline, 2005, p. 94). Changes in student scores due to learning between examination periods can cause changes in statistical values derived from the test. Lastly, items cannot be scored dichotomously since they cannot be subjected to factor analysis. This raises validity concerns with many CTT-based tests since this is exactly how they are coded for they employ multiple choice items (Stevens, 1946).

The necessity of the population’s true scores being normally distributed is the reason why CTT tests are often referred to as norm-referenced. In fact, test items in CTT are included only if they have the property of distributing the population according to a normal curve (Hashway, 1998). This means that students who take such tests are being rank ordered relative to the normal curve that is supposed to represent the scale of human achievement. There have been criticisms that norm-referenced tests are self-fulfilling and can be detrimental to students. If items are omitted from a test because they do not separate students out as desired, then there is no option but for the test to produce a normal distribution of scores. Furthermore,

> [P]upils, seeing themselves labeled in relation to their peers, may limit their own ambitions and some may make their ‘averageness’ come true.
>
> At its best norm-referenced assessment, by letting teachers and pupils see where they stand, may spur them on to higher achievement. At its worst it may demoralize those labeled as being at or near the bottom. Furthermore, it does not set pupils objective standards, so a whole nation could find itself with low achievement levels, simply because it always constructed its own ‘norms’ and never looked outside. It may be better, in the words of the saying, to be a ‘servant in heaven’, rather than ‘master of hell’, but on a norm referenced assessment, the servant in heaven would be on percentile 1, while the master in hell would sit proudly on percentile 100. (Wragg & Wragg, 1997, p. 18).

The quote above illustrates the dangers of being dependent on norm-referenced tests in regards to standards of achievement. Norm-referenced exams usually have a cutoff score (the standard) to determine pass/fail status. If the students are being ranked order against each other on a normal curve, and required to meet a certain population percentage cutoff to pass, that means that a certain proportion of students are guaranteed to fail the exam. This is regardless of how much they may have achieved relative to any standards of learning. To prevent this, CTT tests depend on frequent re-norming of the test to ensure that the test keeps pace with changes in students. For major standardized tests, this could

represent a significant cost to the test developers. Lastly, a study of 440 large scale achievement exams have yielded that none of them actually produced a normal curve leading many to question the assumptions of CTT (Micceri, 1989). To address the inherent flaws of CTT, psychometricians have developed a newer theoretical testing framework that is population distribution independent, item set and number independent, and allows for dichotomous scoring of test items. It is called Item Response Theory (IRT) and will be the major focus of this dissertation.

Item Response Theory

Item Response Theory (IRT) is the predominant theoretical framework for test creation within the current standardized testing movement in the United States, replacing CTT. While IRT does assume the existence of a latent trait that represents achievement and whose magnitude can be made observable as a score on a test like CTT does, the difference is that IRT does not depend on a normal population distribution of scores to meet its assumptions. IRT was originally conceived by Georg Rasch. Rasch developed his models in the 1950's to move away from having to reference the testing population in his own psychometric analyses of reading ability (Rasch, 1960). Rather, he wanted a model that was

> …individual-centered with separate parameters for the items and the examinees… Rasch's point of view marked the transition from population-based classical testing theory, with its emphasis on standardization and randomization, to IRT with it probabilistic modeling of the interaction between an individual item and an individual examinee. (van der Linden & Hambleton, 1997, p. 8).

As noted in the quote above, IRT is based on probabilistic models of students getting an item on a test correctly. To do so requires that items have intrinsic parameters such as difficulty that can be quantitatively determined. One of the claims about IRT is that it is objective in the same way that a ruler measuring the

length of a room is objective. As Benjamin Wright, a friend and fellow colleague of Georg Rasch, writes,

> Objectivity is the requirement that the measures produced by a measurement model be sample-free for the agents (test items) and test-free for the objects (people). Sample-free measurement means "item difficulty estimates are as independent as is statistically possible of whichever persons, and whatever distribution of person abilities, happen to be included in the sample." Test-free measurement means "person ability estimates are as independent as is statistically possible of whichever items, and whatever distribution of item difficulties, happen to be included in the test." In particular, the familiar statistical assumption of a normal (or any known) distribution of model parameters is not required. (Wright & Linacre, 1987, p. 5).

IRT claims to be objective because it uses an external interval-based scale to measure “person ability estimates” (represented by θ), making it independent of the testing population (sample-free). Furthermore, the test items are calibrated to this external θ scale so that regardless of the item sampling on the test, the measures will still be the same for any person (test-free). IRT posits that items and persons can be associated to specific locations on the external scale (Hashway, 1998). This is similar to the idea of using the gradation marks on a ruler to measure length, an analogy commonly used in the IRT literature. The scale of the ruler is independent of the distribution of the lengths of objects it is

used to measure. A person at any point on the θ scale will be able to respond correctly to all items at a lower point on that scale, but not to item farther up the scale (Lord & Novick, 1968).

There are many different models of IRT and the first model that Georg Rasch came up with had only one item parameter: difficulty (b-value). As such, it is referred to as IRT-1PL or the Rasch model and is presented mathematically in **Equation 2.4**. Note that the presented model is actually a popular extension by Allan Birnbaum, who suggested replacing the original normal-ogive model with a logistic model (Birnbaum, 1968).

$$P_i(\theta) = \frac{1}{1 + e^{-(\theta - b_i)}}$$

Equation 2.4 One Parameter Logistic model of Item Response Theory (IRT-1PL)

$P_i(\theta)$ represents the probability of an individual with θ value responding correctly to item *i* with *b* difficulty. In this model, changing the b-value will linearly translate the probability function along the ability or θ scale axis as shown in **Figure 2.1**. The inflection point of the function is at the b-value itself, whose range can extend from negative to positive infinity though in practice is usually between -3 and 3 due to the nature of the logistic curve. Individuals who have the same θ value as the b-value will have a 50% probability of correctly responding to that item.

Since the introduction of IRT-1PL, various other psychometricians have added on to it such that the most popular model of IRT is actually the Three Parameter Logistic model (3PL) which aside from the difficulty parameter (b-

value) has a discrimination parameter (a-value) that determines the slope at the inflection point and a guessing parameter (c-value) that sets the lower asymptote of the probability function. The a-value represents how well an item is able to distinguish between student above and below the item's b-value. The c-value adjusts the probability of responding correctly to an item by accounting for random guessing which is important on multiple choice exams where students could respond correctly based on luck and not ability. The model is presented in **Equation 2.5**.

$$P_i(\theta) = c_i + \frac{(1 - c_i)}{(1 + e^{-a_i(\theta - b_i)})}$$

Equation 2.5 Three Parameter Logistic model of Item Response Theory (IRT-3PL)

The IRT-1PL model is the equivalent to the IRT-3PL when the a-value is one and the c-value is zero. The discrimination parameter (a-value), in theory, can extend from negative infinity to positive infinity, but in actual practice should be a positive value since a negative a-value would have negative discrimination, meaning that the higher the ability (θ) level of a person, the more likely they would respond incorrectly. An a-value of zero would mean that the item does not differentiate at all and that all examinees would have an equal probability of responding correctly. Test makers generally want items with a high a-value. **Figure 2.2** shows the graph of the probability function of different a-values while both b-value and c-value are set to zero. The c-value can range from zero to one. The higher the c-value of an item, the more likely even a low ability person will respond correctly to the item. A c-value of one would mean everyone will

respond correctly to the item regardless of θ value. **Figure 2.3** shows the graph of various c-values when a-value is set to one and b-value is set to zero. The reason IRT-3PL is popular is that it allows for a more nuanced characterization of items. In particular, most standardized exams are heavily dependent on multiple choice items where guessing can play a large role in student scores.

The external θ scale by which examinees and items are measured in IRT is generated during test calibration. Test calibration involves field testing a set of items to a sample of examinees. The resulting response set is dichotomously coded for calibration. The mathematical procedure for test calibration is an iterative two step process first proposed by Birnbaum that involves treating the initial ranking of the examinees on the raw score as if they were the true score while the parameters of each item is individually estimated (1968). Each item must be individually estimated due to the fact that the items are supposed to be independent of each other as agents on a test and their ability to measure students independent of the other items on the test. Then treating the estimated item parameters as if they were the actual item parameters, examinee θ values are estimated. This is also done individually since examinee θ values are independent of each other. The two step process is repeated until the change in estimated values reaches a predetermined threshold considered to be insignificant or when the test calibration reaches a certain number of iterations. A problem with this method of test calibration is that it does not yield a unique metric with a set midpoint or unit of measurement. It is unique only up to a linear transformation similar to how the Celsius, Fahrenheit, and Kelvin scales of

temperature all measure the same phenomenon on different scales and can be transformed interchangeably (Baker, 2001). The person performing the test calibration has to “anchor” the metric based on arbitrary rules. Also note that the θ scale metric

> … depends upon the specific set of items constituting the test and the response of a particular group of examinees to that test. It is not possible to obtain estimates of the examinee’s ability and of the item’s parameters in the true metric of the underlying latent trait. The best we can do is obtain a metric that depends upon a particular combination of examinees and test items. (Baker, 2001 p. 132).

While IRT does not require that the sample of the test calibration examinees to fit any type of distribution, it is important to make sure that the test calibration sample is large and representative of any future possible testing population because the scales generated in test calibration are uniquely a result of the interaction between that examinee sample and the item set sample (Kline, 1993). If all future examinees are to be placed on the θ scale metric based on the original sample of examinees and items, then the test calibration sample should encompass the range of examinees to be tested.

Note how this method of establishing scales is different from how the scales in the physical world are established. The θ scale of the latent trait is dependent on the interaction between item set and the examinee sample to define the nature and scale of the latent trait as a phenomenon. Scales of the

physical world define what the phenomenon of interest is first before the scale is generated such as length, and thus allowing for the conservative conception of one to one measurement between the phenomenon and natural numbers (Savage & Ehrlich, 1992). IRT never defines the latent trait but rather measures what the items in the set have in common relative to the test calibration examinee sample. The definition of the latent trait is normally left up to the test makers and is usually based on the face validity of the items, i.e. what do the items look like they are testing for. All future items are calibrated along the θ scale based on how it behaves when compared to the items that are already calibrated, allowing for the scale to be used with new items and testing populations.

Just as there are limits to the measuring capability of instruments in the physical world, there are limits of measurements for IRT latent trait θ scales. It would be ridiculous to measure the distance between cities in terms of inches by using a ruler since the scale of measurement is so vastly different between the ruler and the distance between cities. Based on this reasoning, it is important that the item set used for test calibration be able to measure the entire range of latent trait values in the test calibration sample. Otherwise, there will be individuals who cannot be measured for they exist outside the range of the latent trait θ scale. In a similar vein, just as there are errors of measurement when using a ruler, there are errors of estimation when using IRT. A ruler is only as accurate to the number and interval of gradations on it. Similarly, a test as a measurement instrument can only estimate an examinee's latent trait θ value as well as the

number of items and the interval between item difficulties allows it to do so. This means that the more items there are and the smaller the intervals of difficulties between items on a test, the better the test will be able to accurately and precisely estimate an examinee's true latent trait θ value.

There are many advantages of IRT over CTT, with the most prominent being that it is independent of the distribution of the *current* testing population due to the external latent trait θ scale. However, IRT does have some drawbacks. We have already mentioned that the latent trait is defined by what the set of items measure in common for the test calibration sample in terms of only the statistics. This commonality is undefined except by the judgment of face validity by test makers. This latent trait is considered to be unidimensional since it is what the items are measuring in common across all students. However, "all [items], written or oral, are a test of language, as well as of the subject matter being taught, so a many-faceted assessment is inescapable." (Wragg & Wragg, 1997, p. 14). It is most likely that IRT items are actually testing for a combination of specific latent traits, some of which are shared across the different latent trait θ scales. However, all are under the guise of one latent trait: the commonality across all items. It would make sense then that all estimated θ values for different latent traits will share some level of correlation with each other. The extent to which each θ value is shared and the amount that is unique to the desired latent trait being measured should be a concern during test construction. This issue refers to the content validity of a test; i.e. is the test actually testing for what it claims?

For the purposes of this dissertation, content validity can be broken into two types: face and logical validity. Regardless of which type is under discussion, content validity cannot be statistically proven, but rather it is a matter of subjective judgment (Allen & Yen, 1979). Face validity is a value judgment made by the test makers and asks if the item looks like it is testing for the desired content on a superficial level. Since most items are submitted by content experts, it is usually assumed that face validity is present (Kane, 2006). Logical validity is used during actual test construction. It depends on a carefully constructed rubric to define different content objectives to be measured and then ensures the logical design of items to meet these objectives by test makers (Allen & Yen, 1979). As was mentioned earlier, IRT is not concerned with issues of content validity since it only measures what items share in common to generate a θ scale for a unidimensional latent trait, and as stated before, the precise definition of the latent trait is left to the test maker's judgment. Consequently, it is important that at every step of test construction and calibration, safeguards are taken to ensure that content validity is still present in the items. This ensures that the determination of the latent trait θ scale will incorporate some measurement for the content that students are expected to achieve on rather than the general latent traits required to take a test. Once test calibration is over and the latent trait established in terms of the statistics with the assumed content validity, the assumption is that calibrating new items to the θ scale should measure only the difficulty of the items on the established latent trait.

IRT is dependent on what is known as parameter and total score invariance. As long as the established values for the parameters of an item stay invariant, the item can be applied to new and different samples of examinees. When these parameters vary, it implies that they are dependent on the sample of examinees used to calibrate them, such that different samples will yield different values of the parameters (Hashway, 1998). This puts the validity of the external latent trait θ scale at risk. A ruler cannot be expected to measure length if its scale is changing based on what it is measuring; an object that is one inch must always be the same length as another object that is also one inch as measured by the ruler. There are a number of threats to item invariance including "context effect, item position effects, *instructional effects*, variable sample sizes, and other sources of item parameter drift that are not formally recognized or controlled for in IRT applications." (Meyers et al, 2009, emphasis added). The fact that instructional effects are considered threats to parameter invariance in the IRT literature indicates that IRT was design to measure static populations in terms of achievement. One could go a step further and say that IRT is measuring for latent traits that are persistent in their values relative to each individual in the population just as height is a persistent characteristic in every full grown adult. This is contrary to the common notion of achievement which is a malleable entity that can and should change over time. If items were sensitive to instructional effects, they would constantly change in their difficulty or b-value year after year depending on the pedagogical practices of teachers and student motivation to learn. Such items would be discarded from the pool of items to be used on an

IRT exam since they violate the invariance principle. Note that parameter invariance refers only to the profile of the population's probability of getting an item correctly based on their θ value and not the distribution of θ values. Total score invariance "implies that the estimate of the position on a latent trait dimension for a particular subject is not a function of the particular subset of [items] drawn from a precalibrated [item] pool used to obtain that estimate." (Hashway, 1998, p. 98). This means that multiple but equivalent versions of an exam should not yield different measurement values for individual students or for students with the same θ value. In keeping with the ruler analogy, different rulers on the same scale should yield the same measurement for the length of an object and that different objects with the same length should have the same measurement value.

IRT is often referred to as criterion-referenced testing. This is because during test calibration, a cutoff θ value of the latent trait could be set based on the achievement level of the initial test calibration sample that was considered acceptable. Any future examinee whose θ value is below this cutoff would be considered as having not met the achievement standards that was originally set. Since the θ scale is independent of the testing population distribution, it is theoretically possible that every single examinee at some future point could reach the cutoff θ value. This is more desirable than norm-referenced where a certain proportion of the testing population must fail the exam because they are in the lower proportion of the current testing population in terms of achievement. However, if you consider the invariance requirement of IRT and the fact that

items sensitive to instruction would be discarded, students already labeled as failing would not be able to change their θ value and therefore can never achieve passing status. In others words, there is an element of pre-determinism when it comes to IRT.

In summary, IRT eliminates many of the flaws of CTT by allowing for items to be graded dichotomously and being population distribution independent. Furthermore, IRT claims to possess a level of objectivity rarely found in psychometrics. For reasons concerned with reliability (the ability to consistently measure students accurately), the invariance principle means that the latent trait must be persistent over time in each individual. Lastly, the validity of latent traits to the actual content we desire to assess is something that IRT does not address, but rather is assumed to be present. Any conclusions drawn from an IRT exam needs to include a thorough examination of the latent trait that students are being measured on. It cannot be assumed that the latent traits represent what the face validity might indicate.

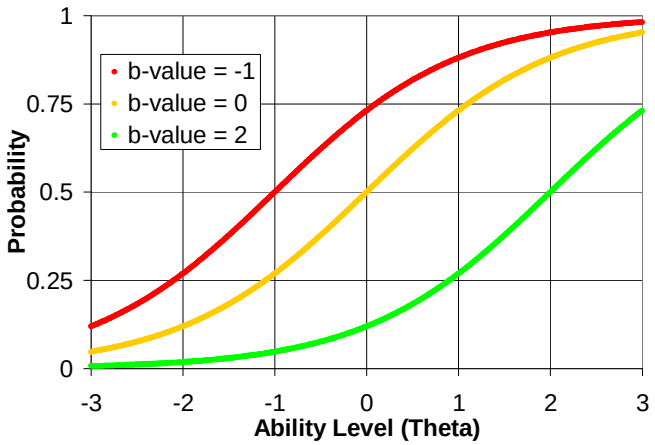

Figure 2.1 IRT-1PL model with varying b-values

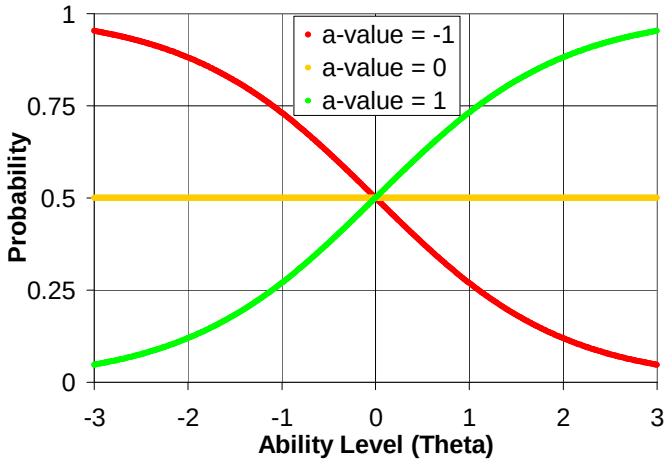

Figure 2.2 IRT-3PL with varying a-values (b-value = 0 & c-value = 0)

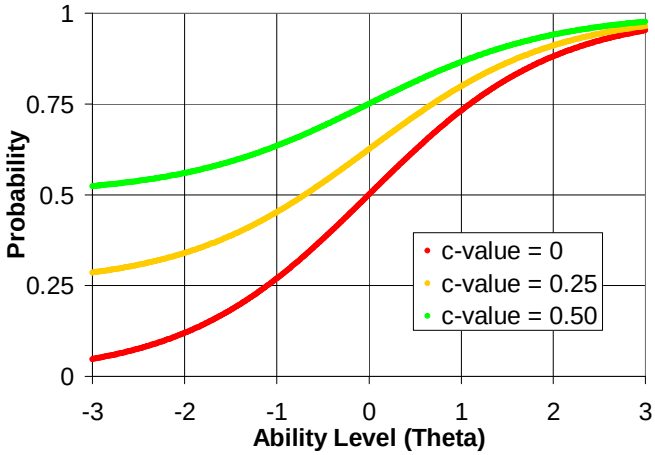

Figure 2.3 IRT-3PL with varying c-values (a-value = 1 & b-value = 0)

The Texas Assessment of Skills and Knowledge Exam

The Texas Assessment of Knowledge and Skills (TAKS) exam is the high stakes summative standardized test administered to students in the state of Texas at the end of each school year*. It was mandated by the state legislature in 1999 in order "to measure the extent to which a [Texas] student has learned and is able to apply the defined knowledge and skills at each tested grade level." (p. 13). The TAKS exam is used to comply with the federal No Child Left Behind (NCLB) Act of 2001 (p. 17).

The exam was developed by Pearson Educational Measurement (PEM) under the guidance of the Texas Education Agency (TEA). The exam will cost the state of Texas at least $279 million between 2005 and 2010 (Pearson Educational Measurement, 2005). There is no time limit on the exam and the final score is determined only by the number of correct responses. This means that there is no penalty for students guessing on the exam (Bobrow, 2004). The TAKS exam is designed and scored based on a Rasch Partial Credit Model (RPCM). For multiple choice and gridded response items that account for the vast majority of TAKS items, this reduces down to the 1 Parameter Logistic Model of IRT (IRT-1PL) (p. 127). There are four domains tested on the TAKS: English Language Arts, Mathematics, Social Studies, and Science. They will be

* Most of the following information comes from the Texas Education Agency's *Technical Digest 2005 - 2006* of the TAKS exam that can be found on their website and so only page numbers will be used to reference which page of the Technical Digest the information can be found. Other references will be cited as usual.

referred in this dissertation hereafter as Reading, Math, History, and Science respectively. Each domain has varying lengths of multiple choice items which are scored dichotomously as correct or incorrect. Furthermore, there may be a few gridded response items that are also scored dichotomously with no partial credit for any work shown. Lastly, the Reading domain contains short response and/or essay items that are scored based on the aforementioned RPCM.

The use of the RPCM was selected by TEA for various reasons, including accommodation of both dichotomously scored (multiple choice and gridded response) and multiple response (short response and essay) type items as well as allowing for the maintenance of a

> one-to-one relationship between derived scores (that is, scale scores) and the raw scores. It is the underlying Rasch scale that facilitates equating of multiple test forms and allows for comparisons of student performance across years. Additionally, the underlying Rasch scale facilitates the critical maintenance of equivalent performance standards across years. (TEA Student Assessment Division, 2006, p. 127).

Furthermore, RPCM allows for TAKS item difficulty estimates (b-values) to be on the same logit scale as student proficiency (θ or ability) estimates.

> Estimates of items being field-tested can be obtained from a form-by-form or a concurrent calibration, with

> the common item set serving as an anchor. In this way, all field-test items can be placed on the same logistic scale as that of the common items.
>
> At the conclusion of these calibrations, all item- and task-difficulty estimates as well as all student proficiency level estimates are directly comparable because they are on the same underlying logistic scale. (TEA Student Assessment Division, 2006, p. 128).

The scores obtained from the RPCM are then rescaled into a TAKS scale score using a linear transformation so that 2100 becomes the cutoff for “Met Standard” performance and 2400 is the cutoff for “Commended” performance. This was done to improve the “readability” of the scores for teachers, parents, and students (p. 129).

Lastly, Reading scale scores were further rescaled based on performance on the essay items. Students must receive at least 2 out of 4 points on the essay items to be considered as having “Met Standard”. Any students who got less than 2 points on the essay section had their Reading scale scores remapped to 2099 automatically if they scored higher on the multiple choice section. This is exactly one point less than the 2100 “Met Standard” performance value. Students must get at least 3 points on the essay items to receive a “Commended” performance designation. Students with less than 3 points on the essay section were rescaled

to 2399 if they scored higher on the multiple choice section. Again this is exactly one point less than the 2400 “Commended” performance cutoff (p. 130).

It should be emphasized that since RPCM reduces to a one parameter model of IRT, it also means that there is a one to one relationship between the raw score and the scale score as mentioned above; for a given raw score there is only one corresponding scale score (p. 127). Thus, a student who responds correctly to all items except for the hardest (highest b-value) item would score the same as a student who responds correctly to all items, but somehow manages to miss the easiest (lowest b-value) item. This obviously does not apply to the Reading domain which gets rescaled based on essay performance.

The TAKS exam is constructed using logical validity to ensure an even spread of the content material to be tested (p. 54). To be included on the exam, items must go through a rigorous control check process first. Items are first generated by item developers under the employment of PEM under the guidelines as set forth by TEA. Item developers are selected for their specific content area expertise and teaching or curriculum development experience. Once the items are submitted to PEM, they are reviewed and then passed on to TEA, who also reviews the new items. The committees at TEA who review these new items are comprised of Texas educators. Texas educators are classroom teachers, curriculum specialists, administrators, and education service staff. Committee members are also selected to be representative of the demographics for the state of Texas. Acceptable items are then field tested, and the results analyzed to ensure fairness across demographics as well as appropriate

difficulty. Once an item is deemed acceptable for inclusion on the TAKS exam, it goes into the item bank along with all relevant statistics such as difficulty estimates to appear on future exams. Since so many content area experts were involved in the creation, review and selection of the items, the items are considered to have content validity (p.150). Recall that there are no statistical tests to show that items or even tests have the desired content validity. To further validate the TAKS exam, correlations to other measures of achievement such as the SAT and the ACT are also done. This allows for the comparison of Texas student to the national data on student performance of these standardized tests (p. 151).

Every live version of the TAKS exam is equated to ensure that comparisons across years are valid. This is done independently by four psychometricians: two from PEM, one from TEA, and one external to PEM and TEA. Equating simply means that the exams are checked to make sure that the logistic θ scale is still the same across test forms and years. If the scale stays the same then comparisons will be valid (p. 156). We used the analogy of a ruler in the IRT section to explain it. If the same ruler is used to measure the length of objects, then it is expected that the same object will always have the same length (with some level of error) regardless of how many times it is measured. Equating then ensures that the same standards are used in all forms of the TAKS exam and is fair to all students.

As shown, a tremendous amount of resources in terms of time, finances, and personnel were put into the development of the TAKS exam. Despite all this,

we will now attempt to show how the TAKS exam has failed our expectations as stakeholders in the educational system and that this failure can be directly attributed to the foundation upon which the TAKS exam was built, IRT-1PL and its assumptions.

CHAPTER 3: Data and Methods

The theoretical premise that the TAKS exam was built upon has already been explored. The remainder of the dissertation will look at the TAKS exam from two different approaches. The first approach uses longitudinal data for real students on the TAKS exam. The second approach uses computer modeling to simulate students taking the TAKS exam. Both approaches yielded different information about the TAKS exam and when taken together, complement each other to tell a complete story for what is happening to Texas students. This chapter begins with a discussion of the origin of the sources of data used in the dissertation as well as the processing that was done on the data to make it suitable for analysis. Then a discussion of the types of analyses done on the data follows as well as why they were necessary in order to determine if and how well the TAKS exam can measure student achievement.

Real World Data

TEA has been maintaining a database that is updated yearly of student performance since the inception of the TAKS exam and is publicly available by mailing TEA's Division of Student Assessment with the request. The database includes all students who have taken the TAKS exam at every grade level and year. Aside from the student performance data, the database also contains information on what schools and districts students come from, as well as other qualitative descriptors such as socioeconomic status, gender, and ethnicity. This provides a wealth of data and nearly limitless ways in which the data can be analyzed. However, any information that could potentially identify specific students was stripped from the database to ensure their privacy as required by law.

Since the research goal of this dissertation is the behavioral trends of how students score on the TAKS exam and the sources of these trends, longitudinal data was the focus. Due to the fact that 2003 was the first year the TAKS was administered, only data starting from 2004 and on would be used to minimize any errors that may have been due to unfamiliarity with the exam by test administrators. In light of the high stakes nature of the junior year TAKS exam as a high school graduation requirement, the dissertation focuses on the longitudinal data for the cohort of 9^{th} graders in 2004 and ended with them in 11^{th} grade in 2006. An immediate concern should be that this eliminated students who failed or dropped out at each grade level. This would cause the data to be biased towards

students who passed each grade level for each year. However, this was a necessary stipulation if the longitudinal trends in students' scores on the TAKS exam were to be investigated.

Data Processing

Once the raw dataset for all students taking the TAKS exam between 2004 and 2006 was obtained from TEA, the dataset had to be processed first before further analysis was possible. Since the raw dataset included many students who did not take the exam in certain years or did not complete all sections, this data was removed so that only students who completed every section of every year were included in the final analysis. This stipulation on the dataset was necessary in order to do a proper longitudinal, inter-domain assessment of IRT. Secondly, students who left more than five items blank in any single section were removed. Since there was no time limit on the TAKS exam nor was there a penalty for guessing, there was no reason then for any student to leave any item blank. An assumption made by IRT-1PL is that all students showed full effort on the exam such that their measured θ value represented an accurate measure of their true ability value. Students who left items blank would result in underestimated ability (θ) values, and so those students were eliminated. It was important that the scale scores and response sets used in the analysis represented students' true effort on the exam in order to accurately estimate item difficulty (*b*-value) and student ability (θ) values.

Students who showed less than full effort was defined as having more than five items blank. This definition was chosen because it represented ~10% of items in each section across years. Students must have invested a substantial amount of time and effort in order to respond to ~90% of the TAKS exam. It was originally suggested that students who have a combined value of incorrect and missing items of more than 25% in each domain be removed. This suggestion grew out of a concern for students guessing on the TAKS. There were four options in the multiple choice items and so even if a student randomly guessed on all items, they should correctly respond to 25% of the items based on chance alone. A 25% cut rate for inclusion into the longitudinal data was ultimately rejected since the TAKS exam was based on IRT-1PL, which did not account for guessing.

There was an inherent danger in processing the data as outlined above. This danger was that the processing would differentially remove more low achieving students than high achieving students. Students of low ability would be more likely to not take the exam or leave difficult items and sections blank. **Figures 3.1** through **3.10** show how processing affected the total distribution at each step. The Raw distributions were the natural distribution of each section of each year before any processing occurred. Note the prominence of students with the minimum scale score indicated by the lone spike to the far left in each graph. These students did not take the TAKS exam and were automatically assigned to the lowest score possible. The Complete distributions represent students who had been processed as stated above and had complete data for that year only.

The distributions do not vary significantly from the Raw distributions† with the exception of removing the lone spike representing students who did not take the exam. The Longitudinal distributions are the set of students who had complete data for all sections of all years from 2004 to 2006. Since these figures are bar graphs and only showed nominal categories of scores, **Figures 3.11** through **3.20** show the scatter plot at the main mode at each level of processing to show any population shift due to processing at scale.

Recall that for the Reading domain, there were short response and/or essay items on the TAKS exam that would cause a further rescaling of the scores beyond what IRT would produce based only on the multiple choice response set. Since essay writing was only assessed in the tenth and eleventh grade, this would account for the discontinuity in the curves seen in **Figures 3.13** and **3.17** at 2099. Also, the Math distributions did not seem to fit one mode, but rather exhibited bimodality for all years. This phenomenon will be explained in the next chapter.

Following processing to produce a suitable Longitudinal dataset to be used for analysis, 139,062 students remained in the dataset. The Longitudinal dataset was then randomly divided into two datasets with N = 69,570 and N = 69,492 respectively. The first dataset (henceforth called the Analytic dataset) was used in the analysis and the second dataset was used as a cross check dataset for the computer modeling and therefore called the Cross Validation dataset. The Analytic dataset was processed one additional step for some of the analyses.

† For a more thorough discussion of the effects of data processing please see Appendix B.

Students who scored perfectly in any domain of any year were removed. The reason for this is that IRT can only measure a student's θ value if the student misses an item. Students who scored perfectly on any section could not be measured by that section; thus any θ values derived for those students were merely a formality to indicate that they scored better than the rest of the population, rather than being an actual measurement. This subset of the Analytic dataset was called the IRT Comparison dataset, since it was only used when comparing IRT measured values in an analysis such as when determining correlation values. The IRT Comparison dataset has an N = 61,311.

Once all processing was done, the dataset had to be converted to a format that coded responses dichotomously as correct or incorrect as dictated in Chapter 13 of TEA's *Technical Digest 2005-2006*, which is necessary for software estimation of student ability (θ) and item difficulty (b) values. Note that even though they should not be treated as incorrect based on IRT; blank responses were coded as incorrect since the assumption is that all students showed their true effort on the exam. Based on students' scale scores on the TAKS exam, this was also how PEM and TEA treated blank responses.

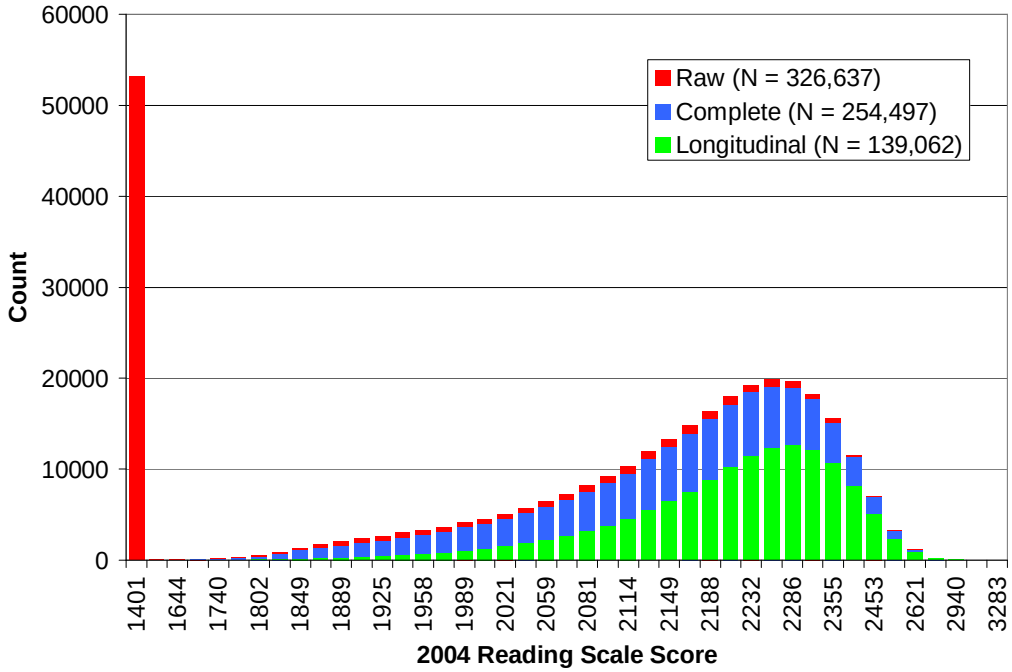

Figure 3.1 Bar graph of the population distribution for 2004 Reading scale scores

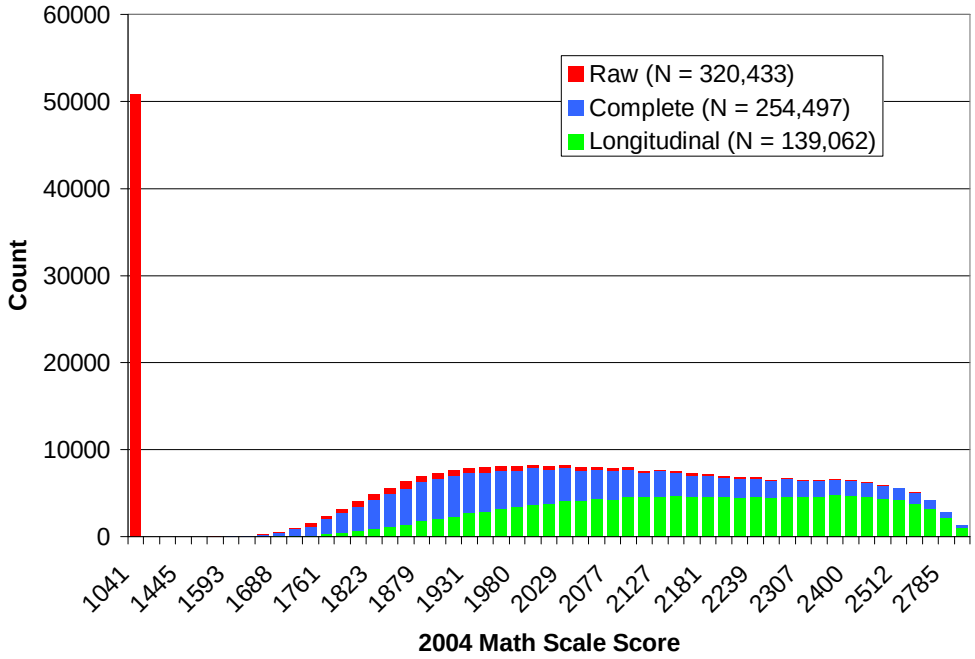

Figure 3.2 Bar graph of the population distribution for 2004 Math scale scores

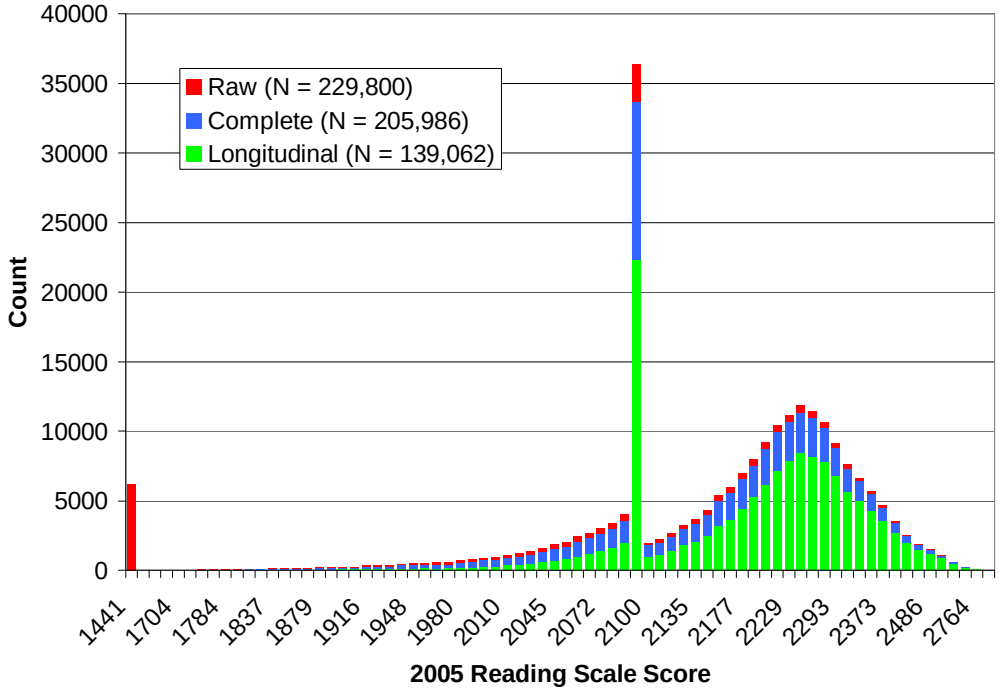

Figure 3.3 Bar graph of the population distribution for 2005 Reading scale scores

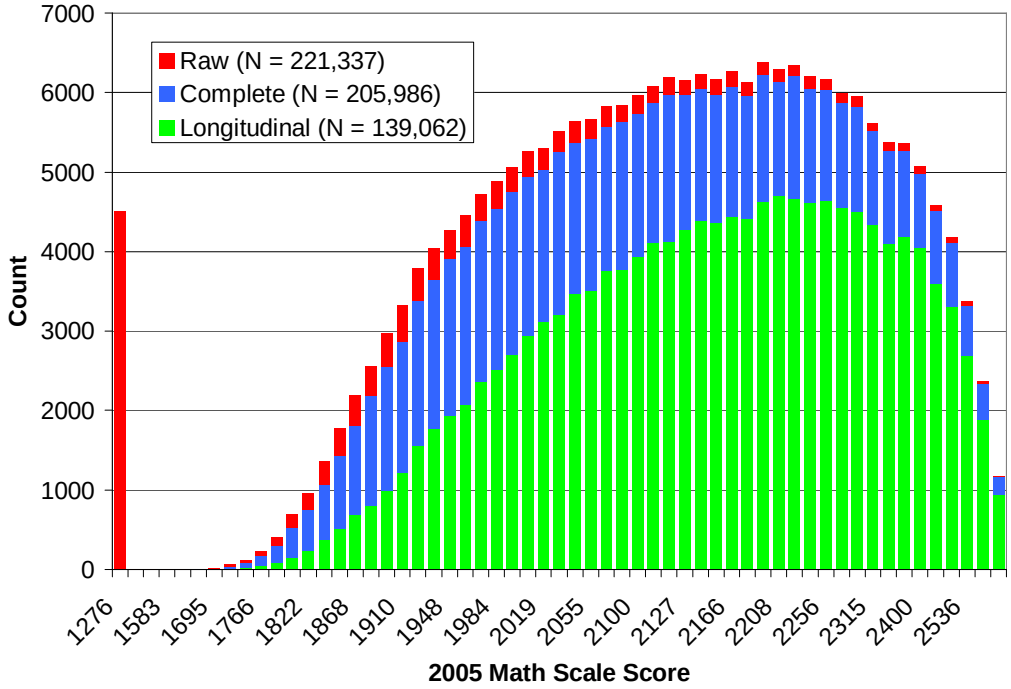

Figure 3.4 Bar graph of the population distribution for 2005 Math scale scores

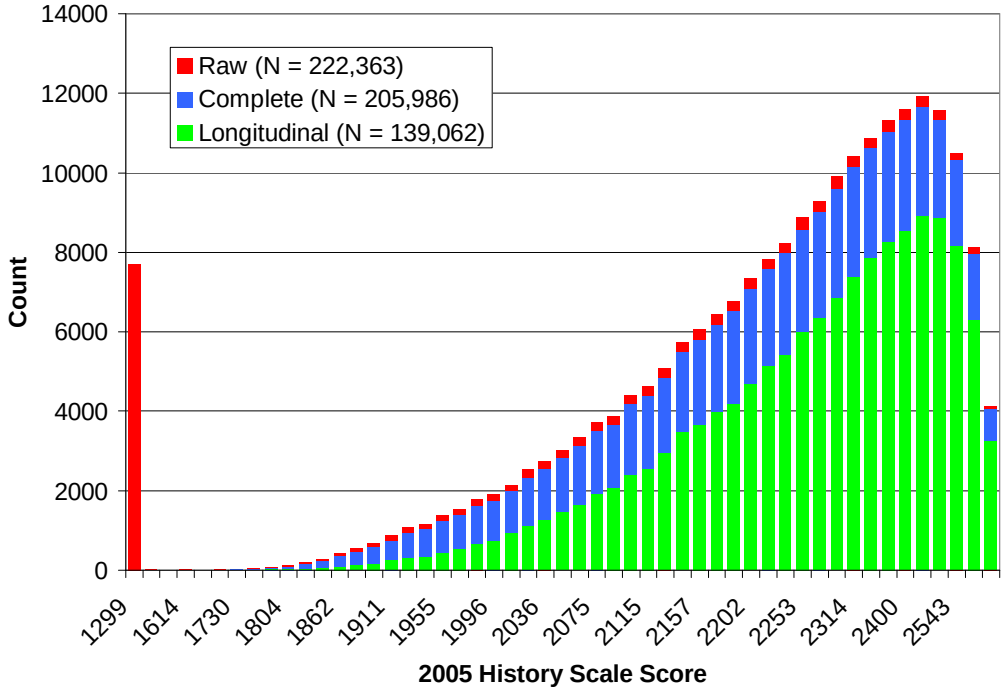

Figure 3.5 Bar graph of the population distribution for 2005 History scale scores

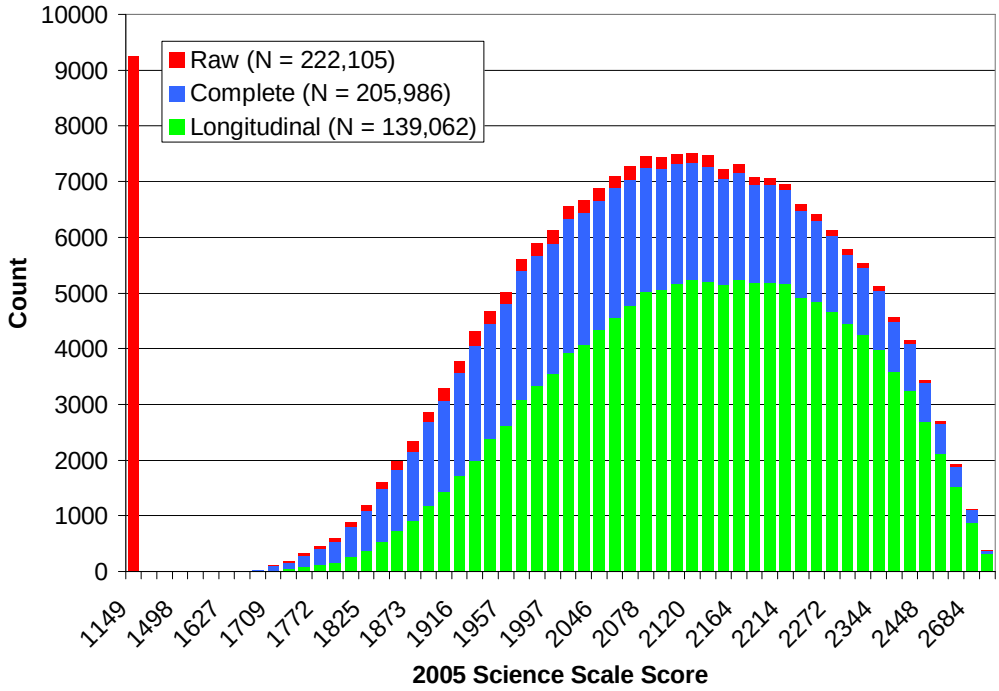

Figure 3.6 Bar graph of the population distribution for 2005 Science scale scores

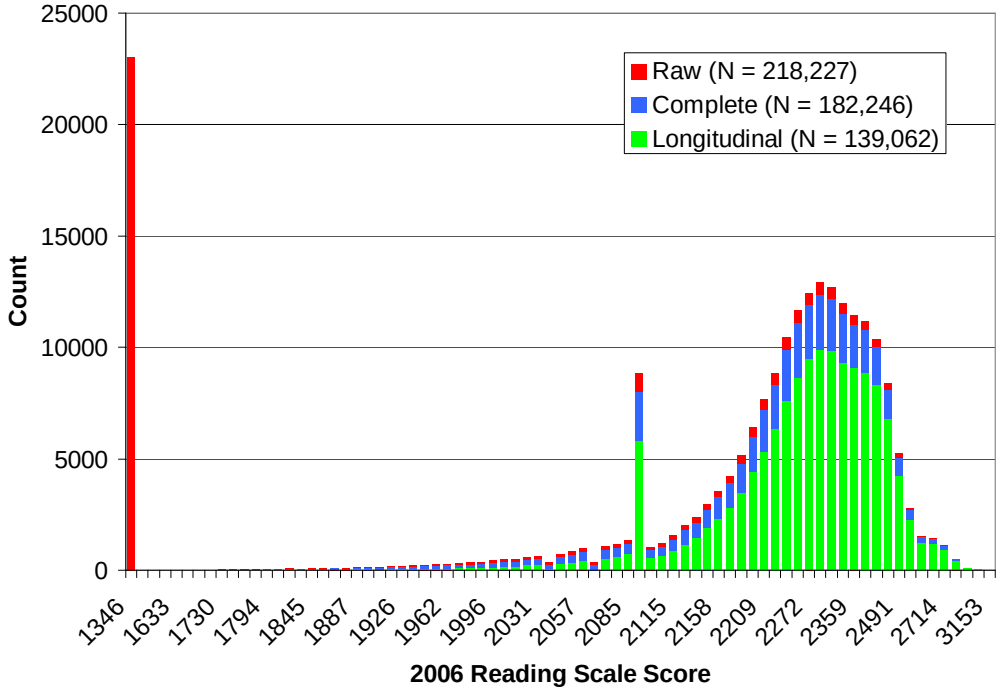

Figure 3.7 Bar graph of the population distribution for 2006 Reading scale scores

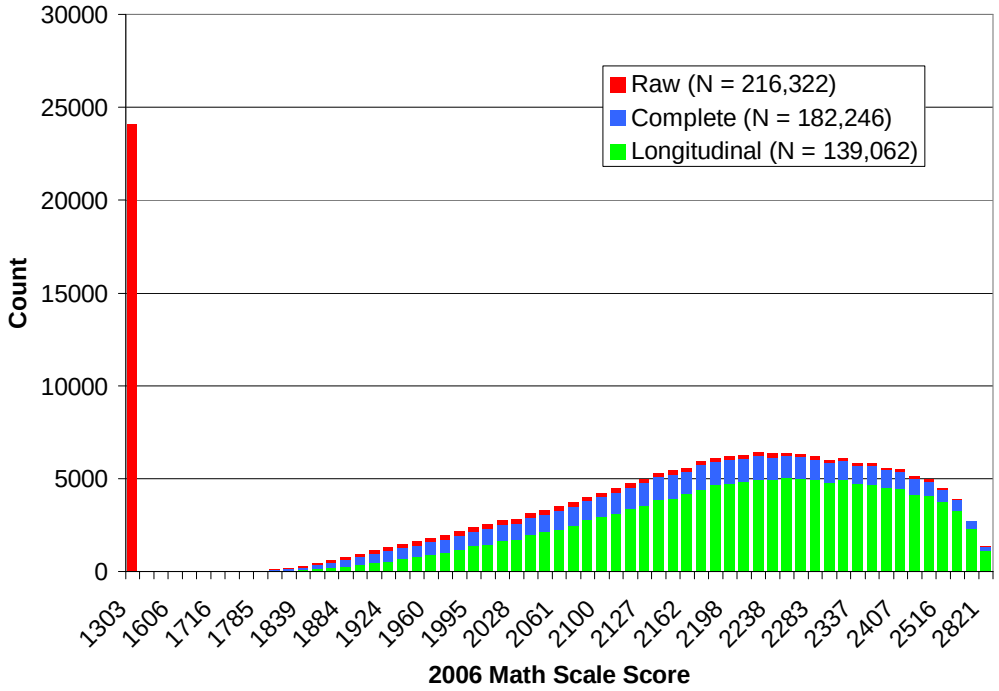

Figure 3.8 Bar graph of the population distribution for 2006 Math scale scores

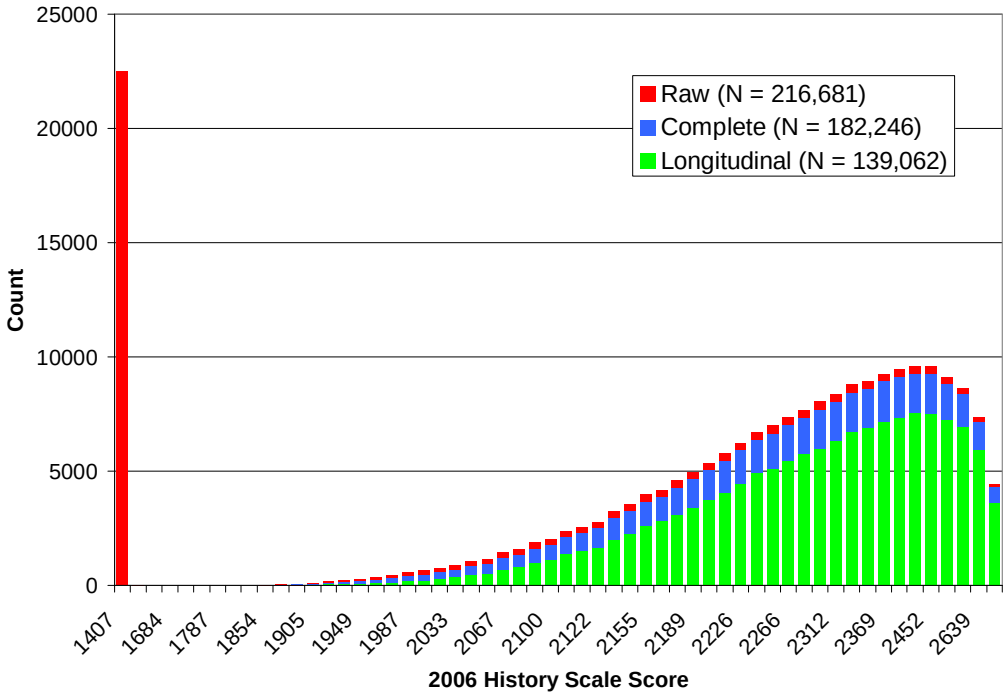

Figure 3.9 Bar graph of the population distribution for 2006 History scale scores

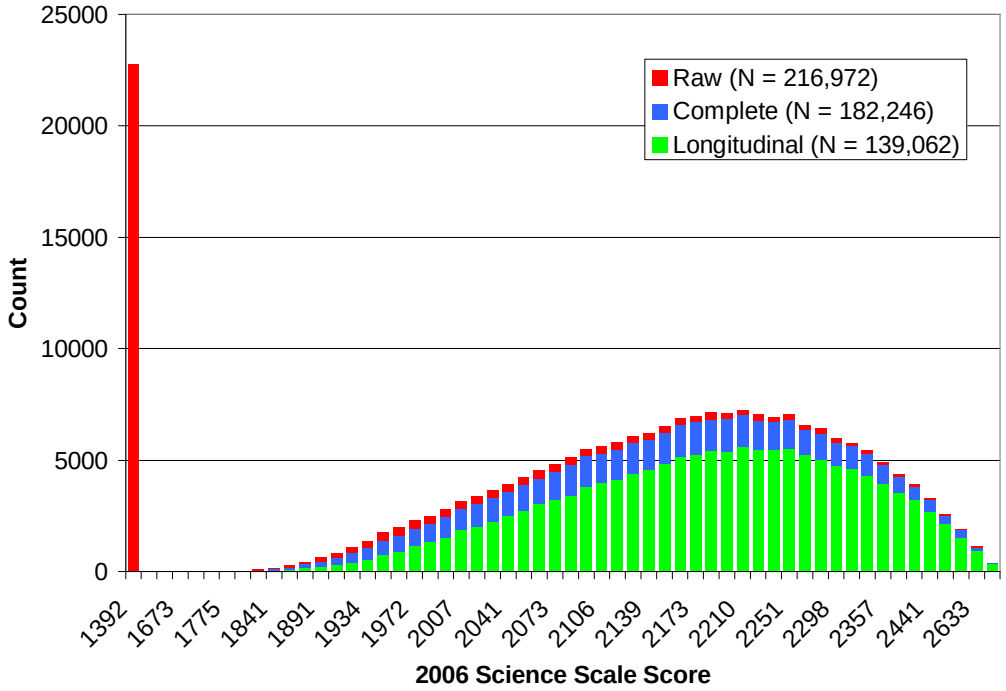

Figure 3.10 Bar graph of the population distribution for 2006 Science scale scores

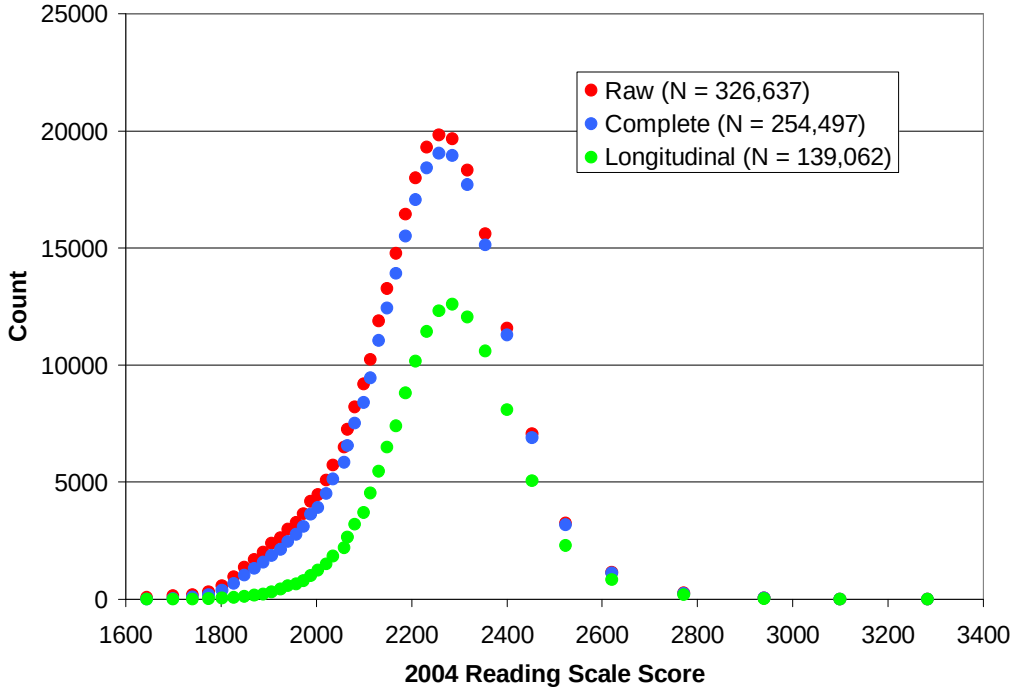

Figure 3.11 Comparison of the distribution of all 3 2004 Reading scale score dataset

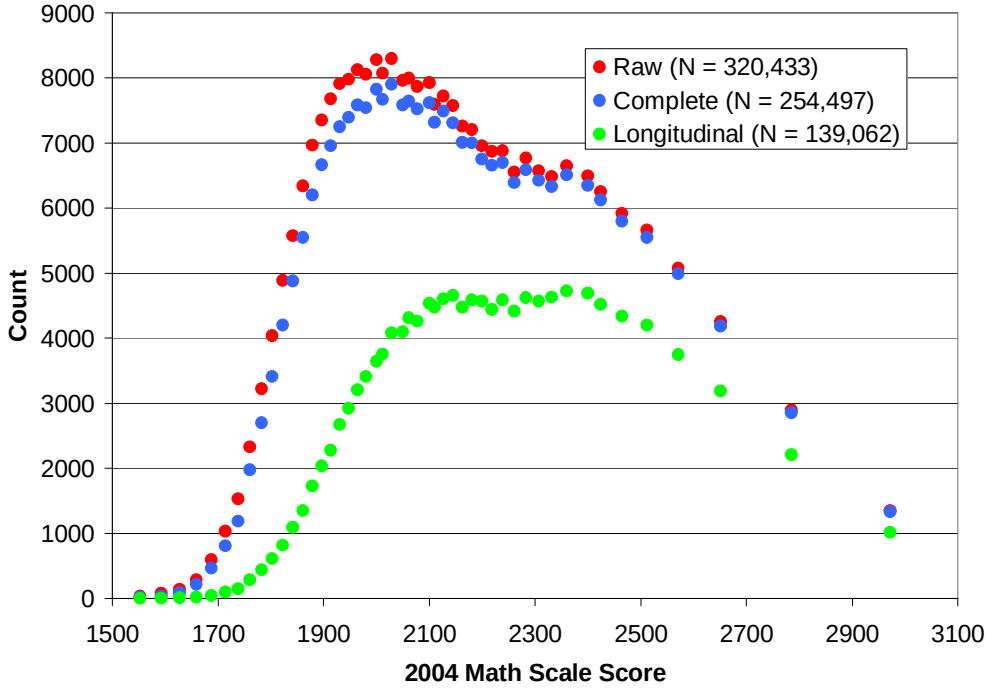

Figure 3.12 Comparison of the distribution of all 3 2004 Math scale score dataset

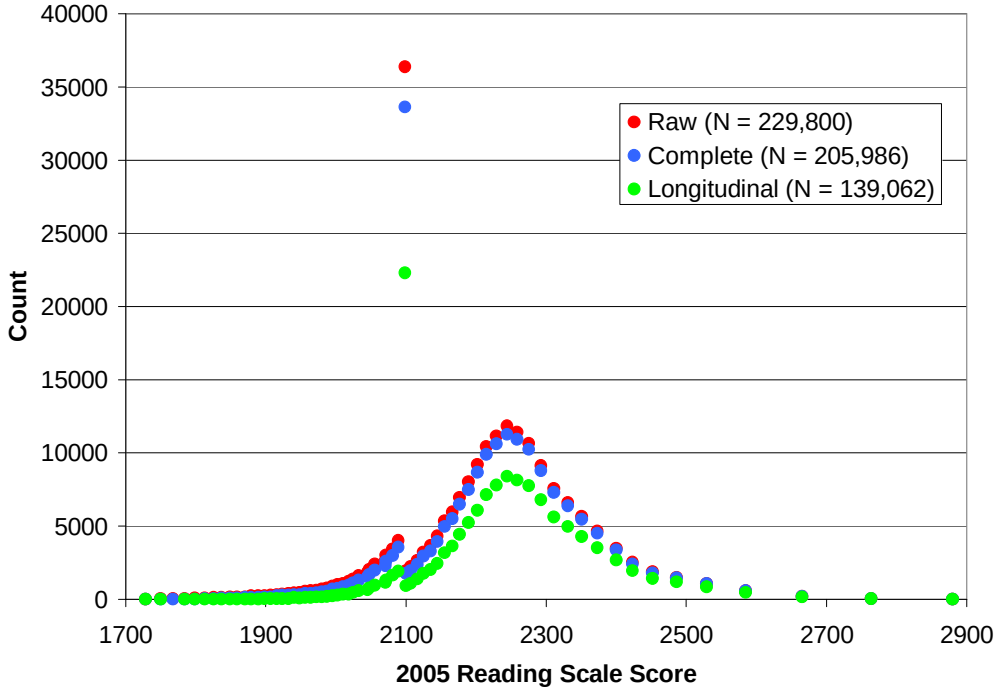

Figure 3.13 Comparison of the distribution of all 3 2005 Reading scale score dataset

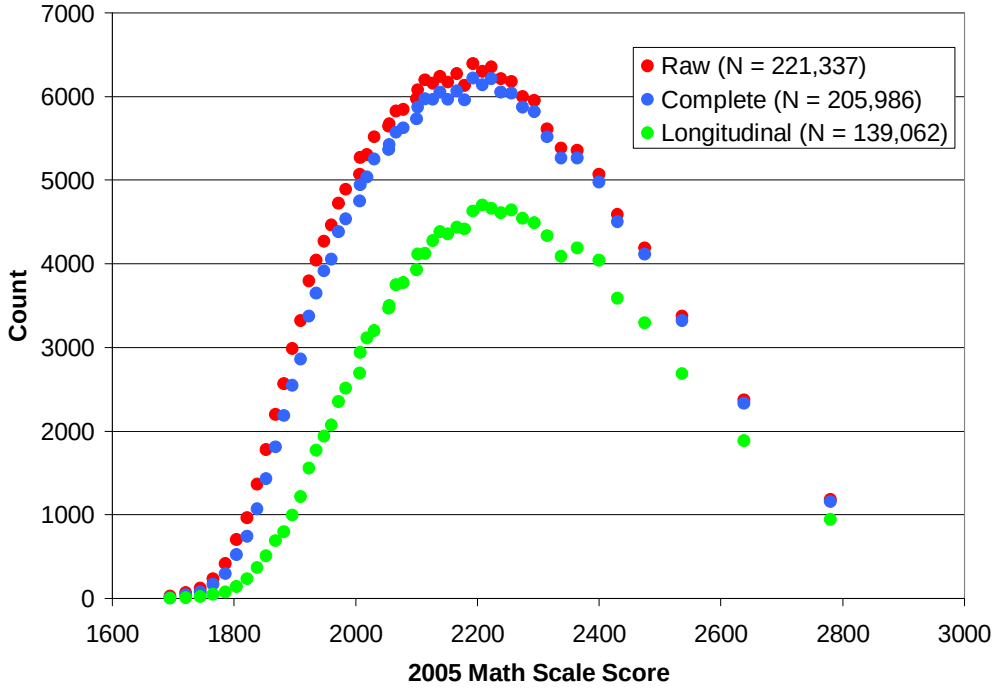

Figure 3.14 Comparison of the distribution of all 3 2005 Math scale score dataset

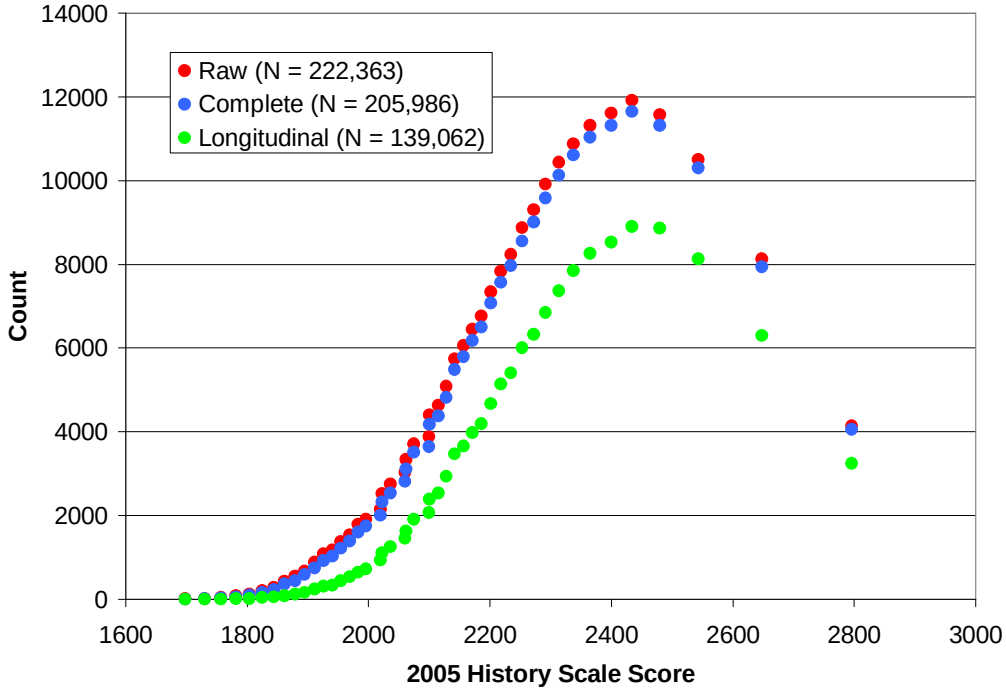

Figure 3.15 Comparison of the distribution of all 3 2005 History scale score dataset

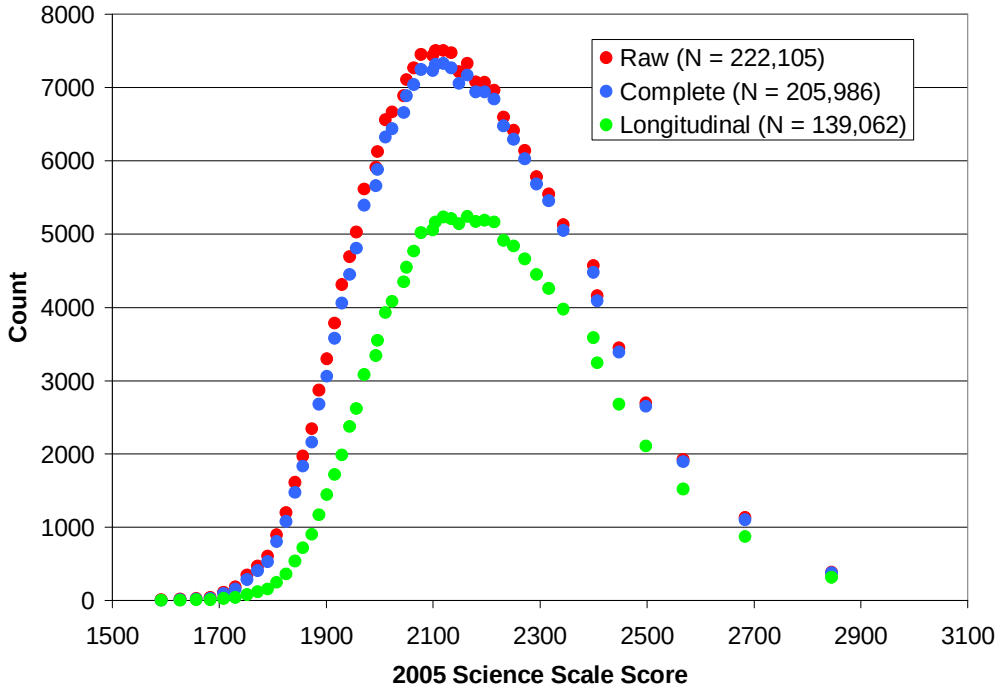

Figure 3.16 Comparison of the distribution of all 3 2005 Science scale score dataset

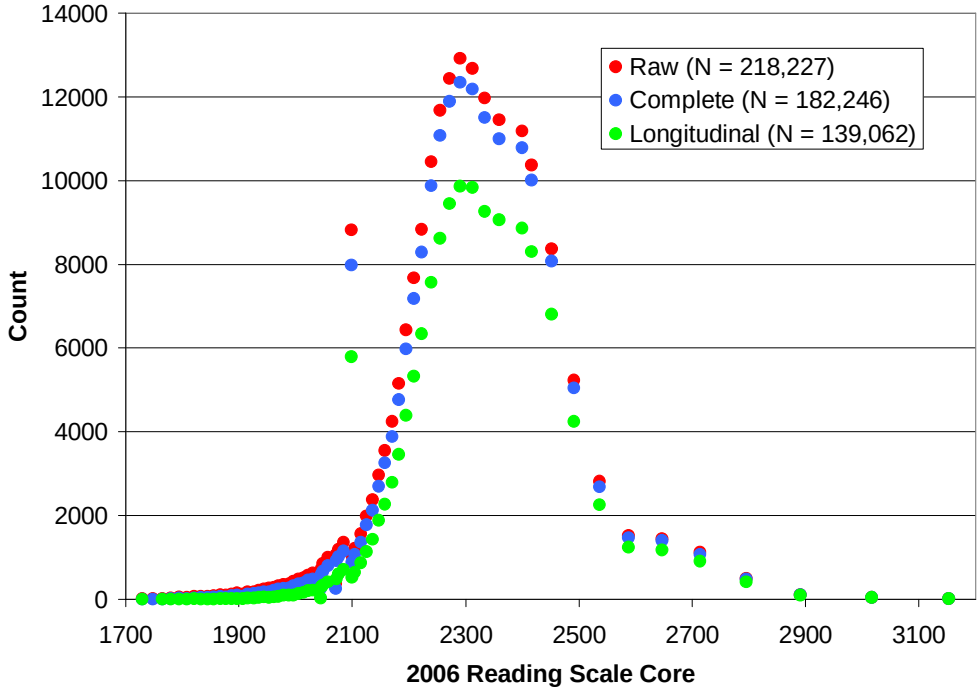

Figure 3.17 Comparison of the distribution of all 3 2006 Reading scale score dataset

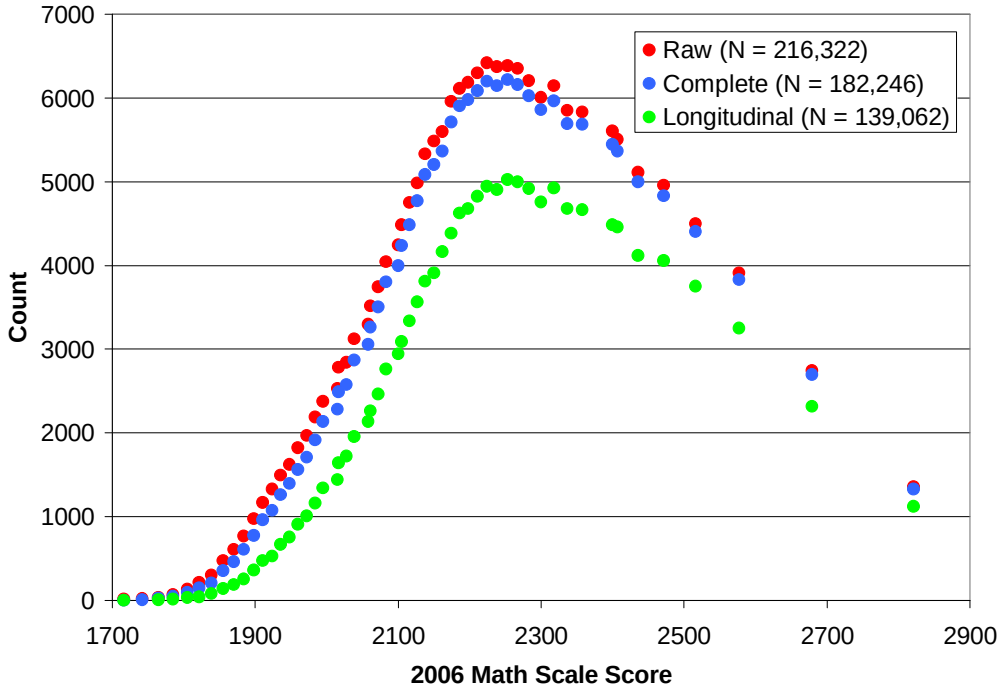

Figure 3.18 Comparison of the distribution of all 3 2006 Math scale score dataset

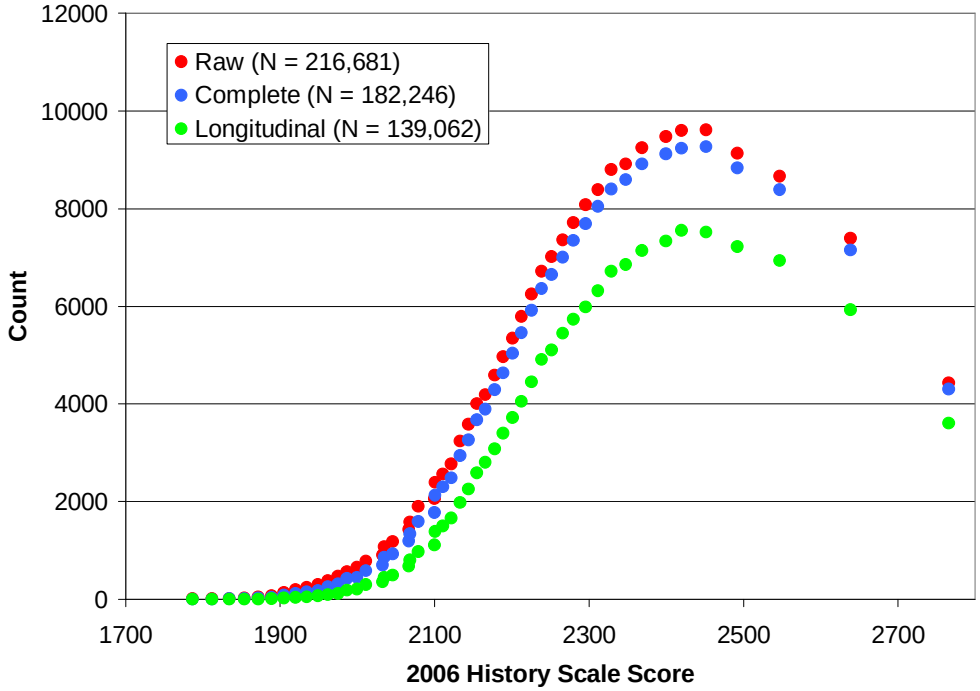

Figure 3.19 Comparison of the distribution of all 3 2006 History scale score dataset

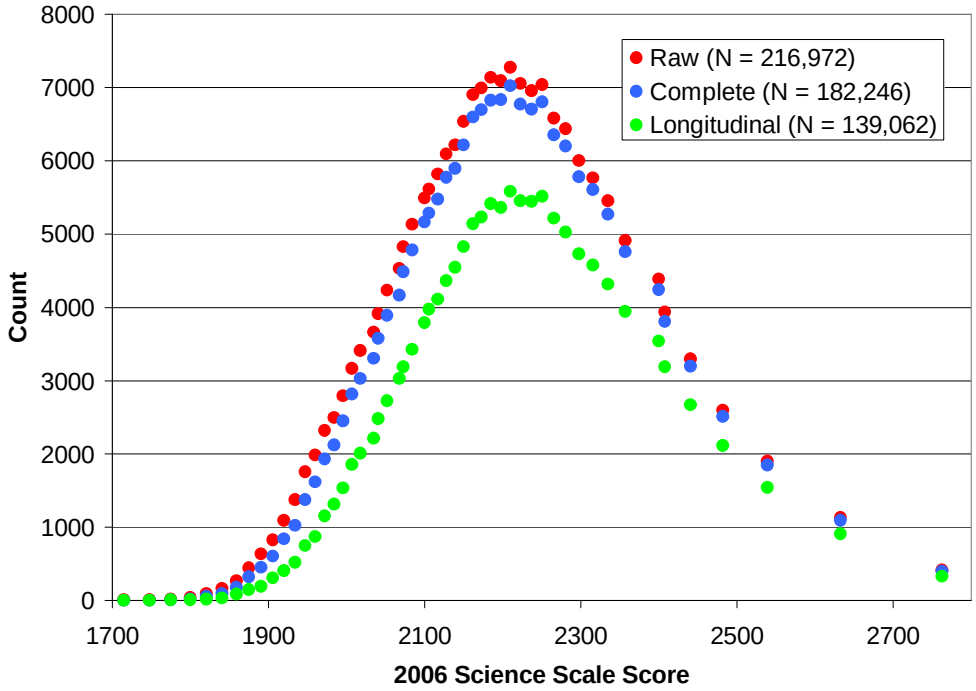

Figure 3.20 Comparison of the distribution of all 3 2006 Science scale score dataset

Computer Model Data

Computer modeling of various complex systems is perhaps one of the most powerful tools made available by modern computer technology. It allows for the exploration of modeled complex systems in ways that are unfeasible in the real world. Parameters can be tweaked or set to run under different conditions to mimic different scenarios. There are many ways to implement computer modeling of complex systems. For the purpose of this dissertation, only one type will be under examination: Agent Based Modeling

In agent based modeling, the computer model contains a number of agents, each of whom will follow a certain set of instructions independently of each other. This allows for the agents to interact with the environment and/or each other. The agents can represent anything that exists as discrete entities in a population such as people, atoms, or trees. Agent based modeling is useful to show complex behavioral interactions and emergent properties that would not normally be seen otherwise. This makes agent based modeling uniquely suited for studying population dynamics within a set system. For this dissertation, rather than modeling a complex biological phenomenon, student behavior on the TAKS exam is modeled. NetlLogo is the modeling platform used due to the author's familiarity with it. NetLogo belongs to the family of StarLogo coding languages that has classically been used in agent based modeling. StarLogo coding languages refer to agents as "turtles" by tradition and so agents will sometimes be referred to as such in this dissertation.

Model Design

A computer model was generated to test the theoretical IRT framework of the TAKS exam. The computer model contained 30,000 turtles (the students) whom each possessed five latent traits: profile (LTP), reading (LTR), math (LTM), history (LTH), and science (LTS). LTP was included as an all purpose latent trait to link the domain latent traits to each other and will be explored in greater detail in **Chapters 4** and **5**. It was assumed that each domain latent trait (LTR, LTM, LTH, and LTS) was responsible for the scores within that domain for all years. This may not be the case at all since the scales were generated independently at the domain and grade level on the TAKS exam. However, the assumption was reasonable if the TAKS exam has content validity. It could be argued that each TAKS exam for year contained specific objectives entailing different knowledge and skills that were not tested any other year. While this could be true and the analysis did take this under consideration, the analysis of both real world data and computer model data showed that this was not the case at all.

Drawing from the diffusion of innovation literature which makes parallels to the epidemiological literature, the computer model used a “disease” type simulation of generate latent trait values. One of the turtles would randomly catch a latent trait “disease”. The “disease” was then spread by physical contact as the turtles moved randomly about the map until all turtles are infected. The time of infection for each turtle was recorded and used to determine the value of the latent trait. The longer it took for a turtle to catch the latent trait “disease” the

higher the value of their latent trait. This method of distributing the latent trait values was more naturalistic based on the diffusion of innovation and epidemiological literature than randomly assigning latent trait values to the turtles (Rogers, 1995 and Boslaugh, 2007). Note that each latent trait had to be generated independently to ensure orthogonality between the domains. Simultaneous generation of the latent trait values led to localized population interaction that caused the domains to be correlated to each other. Since IRT claims to be population distribution independent, how the latent trait values were assigned should not matter affect the running of the IRT model.

Once the latent trait values were determined, a linkage value could be set for each of the domain latent traits to LTP based on a linear model of shared variance as follows:

$$LTX_{new} = (1 - Linkage) * (LTX_{original}) + (Linkage) * (LTP)$$

Equation 3.1 Linkage Equation to Link Domains to the Profiling Latent Trait

The Linkage value is a measure of how related the domains were on the TAKS exam. LTX represented any of the domain latent traits with $LTX_{original}$ representing the value obtained from the initial “disease” simulation. This allowed the domain latent traits to have a proportion of the variance be orthogonal to each other and the rest to be shared with each other. Users of the computer model would be able to set how related the domains were to each other in order to model how IRT-1PL would work under different conditions. Lastly, the turtle population could have each domain latent trait scaled to any mean and standard deviation value desired. This allowed for the modeling of populations with

different overall ability levels relative to the TAKS exam as well as scaling the different domains relative to each other.

Thus far the model has only been concerned with setting up the initial population conditions. Once the initial population parameters were set, the turtles were then tested on the TAKS exam using the item b-values estimated from the Analytic dataset for each domain and year of the TAKS exam. The testing procedure inserted the turtle's latent trait value and the item's b-value into the IRT-1PL equation (**Equation 2.4**) to determine the probability that the turtle will respond correctly to the item. This probability value was then compared to a value from a random number generator. If the random number generator's value was less than the determined probability of responding correctly to the item, the turtle received a correct response on the item and if the random number generator's value was greater than the probability, the turtle received an incorrect response on the item. This generated a response set for each turtle for each domain and year. The response set was then inputted into the estimation software PARAM-1PL along with previously determined b-values of the items and a resulting turtle θ estimated for each domain.

The value of this model was that since each latent trait value for each turtle was established *a priori*, it allowed for the comparison and evaluation of IRT based exams back to the original latent trait values. This is impossible to do in real life since the true value of the latent traits could never be known. Furthermore, the model represented a perfect world scenario in terms of the execution of an IRT based exam. The turtles did not get tired or hungry or bored.

They did not leave answers blank or guessed on any item. Thus, the turtle θ values obtained from the model represented their true effort with any difference from the turtle's declared domain latent trait value due to random measurement error.

Analyses

Recall that the dataset that was available from TEA contained so much data beyond just the TAKS scores that the types of analysis that could be done were virtually limitless. However, within the designated goal of this dissertation, there were only a few types of analysis that needed to be done to understand the psychometric behavior of the TAKS exam. The analyses were broken down into three general categories: student behavioral trends, item behavioral trends, and the interaction between the students and the TAKS exam.

Longitudinal data for TAKS scores was used in the student behavioral trends to analyze how the scores changed. An examination of how students were behaving across years and domains was important since the scores are the end results of IRT-1PL. Knowing the end outcome would allow us to understand the mechanistic principles that govern how IRT-1PL arrived at student scores as well as the role the students played. On the other hand, item behavioral trends allowed us to see how the items were related to each other across years and domains. To do this, we mimicked how IRT-1PL analyzed items by generating Item Response Functions and compared them against each other. Lastly, we needed to see the interaction of the students with the different domain on the TAKS exam across years to understand from a top down perspective of what was happening each time the TAKS exam was administered. We used structural equation modeling (SEM) to accomplish this analysis. SEM focused on the TAKS exam as a whole, incorporating the interactions between students and domains

across years. This allowed us to see how the variances in student scores in each domain were shared, allowing for the elucidation of the relationships between the different domains on the TAKS exam.

Student Behavioral Trends

The student behavioral trends analysis looked at how students were behaving in terms of their TAKS scores. Of all the domains, only Math and Reading were tested for all three years. Since Reading got rescaled from the IRT-1PL score obtained on the multiple choice section to reflect the essay component, the analysis focused only on Math. The analysis started off by looking at how students were doing from one year to the next in terms of their scores. This showed a general trend based only on the actual scores of how students performed from year to year. Then to remove any dependence on score values, only the changes in scores were examined for consecutive years. Thus, the difference between the third and second year were compared to the difference between the second and first year. This indicated how the score change one year affected the score change the next year. Persistent trends in the longitudinal data would indicate a systemic cause.

When looking at longitudinal TAKS data, there were two phenomena that we needed to be aware of: ceiling and/or floor effects and regressions to the mean (RTM) (Taris, 2000). The TAKS exam measured students on a finite range of scores; that is, there were upper and lower limits on the scores. So the higher a student scored, the less room they had to improve since they were reaching

the ceiling. Rather, they had more room to decrease. When a student achieved the highest possible score, there was no where else to go but down on the TAKS exam (Byth & Cox, 2005). This is the ceiling effect. The same would be true for students scoring in the lower extreme of the range of scores and is called the floor effect. Since none of the students in the cohort under study were close to missing all items on any section, there should have been no floor effect in the dataset, but considering that many students received a perfect score on one or more section, there should be a ceiling effect. RTM was based on the idea that every student has a true score and that any measurement made of this true score would entail some level of random measurement error. Thus, the more a student scored above their true score just by chance alone, the more likely they would score closer to their true θ value the next time. Since RTM is due to random error, a student's score should be normally distributed about the true θ value. Furthermore, the changes in the population of student scores from year to year would also be normally distributed. Some students would be lucky and score higher than they should, and some would be unlucky and score lower, but most would score close to their mean. When RTM is combined with a ceiling effect, the distribution should become non-normal. Students who scored at the limits of the TAKS exam would not have the same standard error of measurement as the students who scored close to the mean. Instead, the distribution should be logistic, with more students residing in the end tails of the distribution. Population distributions were be tested using Q-Q plots which graphed the actual distribution against the desired theoretical distribution.

Item Behavioral Trends

There were many ways in which the TAKS items could be analyzed. In this dissertation item behavioral trends analysis was only related to how IRT analyzed items: by determining an Item Response Function (IRF). An IRF was determined by looking at the relative rank ordering of the students on the exam and determining the ratio of students at each ranking that correctly responded to the item of interest. The use of a ratio is the reason why IRT is population independent. By only looking at the ratio of correct response at each ranking in the dataset for an item, the actual distribution of the underlying population became irrelevant. Item analysis was done to determine how items behaved when compared against the different domains and years. If domains tested for unique knowledge and skills (ergo are orthogonal), using the wrong domain would not generate a clear IRF. Instead, the IRF would be a straight horizontal line across the plot due to the fact that the probability of a correct response is the same regardless of the ranking. If items exhibited the same IRFs across domain scales, then the scores in each domain did not reflect the measurement of a unique latent trait. This would compromise the use of the TAKS as an instrument for accountability as required by NCLB.

Student and TAKS Exam Interaction

The last type of analysis used in this dissertation looked at the interaction between students and test to understand the relationship between each other. Structural Equation Modeling (SEM) is a statistical technique classified as Causal Modeling or Path Analysis. Unlike most statistical tests, SEM allows for the evaluation of complex predictive (causal) relationships based on correlation, which has stirred up controversy among statisticians since one of the axioms of statistics is that "Correlation does not prove causality." We left the argument of SEM for others to debate (Anderson & Gerbing, 1988; Kelloway, 1995; Mueller, 1997; Chambers, 2000) and said that SEM allowed us to estimate the amount of shared variances and correlations among latent variables which was necessary to our analyses. In SEM, a causal model is generated based on a sound theoretical premise. Data is collected and the causal model tested to see if the data fitted the model by looking at the covariance matrices and determining shared variances among variables. Under normal circumstances, causal models are constructed prior to data collection. However, in reality, many SEM analyses use existing data derived from various databases.

Model fit cand be assessed via different indices for goodness-of-fit. The most common one is the χ^2 test. Normally, in a χ^2 test, a large χ^2 value would indicate a significant difference between the expected and observed values. However, in SEM, the χ^2 value should indicate an insignificant difference between the predicted and actual covariance matrices if the proposed model is to

be accepted. There is a caveat though, and that is that as sample size grows, the power of the statistical test underlying SEM also grows. Robert Ho wrote:

> With a great deal of statistical power, almost every reasonable model will be rejected if only the chi-square value and the associated probability are considered. Therefore, given… large samples, a proposed model can easily fail to fit the data statistically, even though the discrepancy between the sample covariance matrix and that reproduced by the parameter estimates of the proposed model may be insignificant from a practical point of view. Given these limitations, the researcher should complement the chi-square measure with other goodness of fit measures. (Ho, 2006 p. 285)

According to Chen et al. (2007), a SEM sample size of less than a hundred is considered small, while one to two hundred is medium, and anything over two hundred is large. Since all of the SEM analyses done in this dissertation had N values of well over 20,000, we should also rely on other measures for goodness of fit. The Root Mean Square Error of Approximation (RMSEA) is another popular index for goodness-of-fit between the proposed causal model and the data. RMSEA is a measure of discrepancy per degree of freedom, and asks the question “How well would the model, with unknown but optimally chosen values, fit the population covariance matrix if it were available?” (Browne & Cudeck,

1993, pp. 137 -138, as quoted in Ho, 2006, p. 285). Values from 0.05 to 0.08 are considered acceptable, 0.08 to 0.10 are mediocre, and anything above 0.10 is a poor fit (Ho, 2006). The last index of goodness-of-fit that researchers usually looked at is the incremental fit measures. These indexes compared the fit of the independence model and the proposed model. The independence model assumes that none of the variables are correlated with each other. In general, values greater than 0.90 are considered good.

The utility of SEM is that it allows for the determination of latent traits that are not directly measured but rather are calculated from the shared variance of variables that were actually measured. As such, the latent traits incorporate very little measurement error since random measurement error should not contribute to shared variance. However, just like in IRT, the latent traits are not truly defined except as the shared variance between measured variables. It is up to the researcher to give the latent traits a definition based on a theoretical premise.

CHAPTER 4: Real World TAKS Exam Data Analyses

Research must be grounded in real world data if it is to have a practical influence in people's lives as opposed to dealing with only the theoretical modeling. This chapter will deal with the analysis of real world data to show the trends that were occurring among real students in the state of Texas between 2004 and 2006. Knowing the trends is the first step in understanding how the TAKS exam and implicitly how the psychometric model of IRT-1PL works. Many of the trends were to be expected at the surface level, though some were surprising and revealed how the assumptions made by IRT-1PL affected student scores.

Software Validation

The freeware program PARAM-1PL can be used to estimate student θ and/or *b*-values from student response sets via maximum likelihood estimation (MLE) (Rudner, 2007). It was important to determine if the program was able to estimate values that were similar to those determined by TEA and PEM. Using only students' response sets to the multiple choice items, PARAM-1PL was asked to estimate both student θ and *b*-values simultaneously for the Analytic dataset. **Table 4.1** shows a correlation matrix of the resulting estimated θ and the student's scale scores from the IRT Comparison dataset (note the change in datasets). Recall that the IRT Comparison dataset was a subset of the Analytic dataset with students who scored perfectly in any domain removed. This dataset was used since the interest laid in determining the ability of PARAM-1PL to estimate θ values and to be able to compare them to the TAKS scale scores. Since students who scored perfectly were not actually measured by either TEA or PARAM-1PL, it was reasonable to remove them.

Aside from the Reading domain, the PARAM-1PL estimated θ values were perfectly correlated to the scale scores. The Reading domain contained short response and/or essay items that caused a further rescaling of student scale scores as mentioned before, so it was not surprising that it was the only domain that did not have perfect correlations between the estimated θ and scale scores. This served to validate PARAM-1PL as parameter estimation software that was very similar to that used by PEM and TEA.

Furthermore, it also proved that the TAKS exam met the invariance principle of IRT. Simultaneous estimation of student θ and b-values was the equivalent of treating the students in the dataset as if they were the test calibration sample. The fact that the interaction between the set of items and students under analysis would yield scales that were in alignment with those determined during test calibration after several years was astonishing in terms of the persistence and invariance of the latent traits that the TAKS exam was measuring.

	R04 SSC	R04 Est. θ	R05 SSC	R05 Est. θ	R06 SSC	R06 Est. θ	M04 SSC	M04 Est. θ	M05 SSC	M05 Est. θ	M06 SSC	M06 Est. θ	H05 SSC	H05 Est. θ	H06 SSC	H06 Est. θ	S05 SSC	S05 Est. θ	S06 SSC	S06 Est. θ
R04 SSC	1.00																			
R04 Est. θ	**0.97**	1.00																		
R05 SSC	0.58	0.55	1.00																	
R05 Est. θ	0.63	0.62	**0.73**	1.00																
R06 SSC	0.59	0.56	0.60	0.58	1.00															
R06 Est. θ	0.61	0.59	0.54	0.62	**0.79**	1.00														
M04 SSC	0.57	0.55	0.50	0.55	0.50	0.54	1.00													
M04 Est. θ	0.57	0.55	0.50	0.56	0.50	0.54	**1.00**	1.00												
M05 SSC	0.56	0.54	0.52	0.57	0.52	0.55	0.82	0.82	1.00											
M05 Est. θ	0.56	0.54	0.52	0.57	0.52	0.55	0.82	0.82	**1.00**	1.00										
M06 SSC	0.52	0.50	0.48	0.52	0.51	0.54	0.77	0.77	0.81	0.81	1.00									
M06 Est. θ	0.52	0.50	0.48	0.52	0.51	0.54	0.77	0.77	0.81	0.81	**1.00**	1.00								
H05 SSC	0.62	0.61	0.54	0.63	0.54	0.58	0.64	0.64	0.68	0.68	0.61	0.61	1.00							
H05 Est. θ	0.62	0.61	0.54	0.63	0.54	0.58	0.64	0.64	0.68	0.68	0.61	0.61	**1.00**	1.00						
H06 SSC	0.56	0.56	0.47	0.57	0.49	0.54	0.58	0.58	0.60	0.60	0.60	0.60	0.74	0.74	1.00					
H06 Est. θ	0.56	0.56	0.47	0.57	0.49	0.54	0.58	0.58	0.60	0.60	0.60	0.60	0.74	0.74	**1.00**	1.00				
S05 SSC	0.59	0.57	0.53	0.60	0.52	0.56	0.72	0.72	0.76	0.76	0.70	0.70	0.75	0.75	0.69	0.69	1.00			
S05 Est. θ	0.59	0.57	0.53	0.60	0.52	0.56	0.72	0.72	0.76	0.76	0.70	0.70	0.75	0.75	0.69	0.69	**1.00**	1.00		
S06 SSC	0.55	0.54	0.48	0.56	0.50	0.55	0.68	0.68	0.71	0.71	0.73	0.73	0.70	0.70	0.73	0.74	0.77	0.77	1.00	
S06 Est. θ	0.55	0.54	0.48	0.57	0.51	0.55	0.68	0.68	0.71	0.71	0.73	0.73	0.70	0.70	0.74	0.74	0.77	0.77	**1.00**	1.00

Table 4.1 Correlation matrix of scale scores and estimated θ to validate PARAM-1PL using the IRT Comparison dataset (N = 61,311)

Student θ and Item b-value Determination

IRT-1PL is based on the idea that each student possesses an ability (θ) latent trait that determines the probability of that student getting an item of *b*-value difficulty correct. Thus, it is important to be able to determine the θ values of the students taking the TAKS exam as well as the *b*-value of each item. Student θ values on the TAKS are reported as a scale score based on the following linear transformation:

$$\text{Scale Score} = (\theta * T1) + T2$$

Equation 4.1 Linear Transformation of θ to Scale Score

T1 and T2 are the constants that allow for the scaling of scores such that 2100 is the "Met Standard" performance cutoff and 2400 is the "Commended" performance cutoff. The values of these constants are reproduced from TEA's *Technical Digest 2005-2006* (p. 131) on **Table 4.2**. Note that these values were established when the TAKS was first field tested and is the same regardless of year since the θ scales are supposed to be static. It is possible then to reverse the transformation back to the original θ values as determined by TEA using these values and **Equation 4.1**. **Table 4.3** shows the student θ value descriptive statistics for each domain and year as determined by TEA using the Analytic dataset. These student θ values will be called TAKS θ since they were the values used by TEA to determine the TAKS scale scores. It is now possible to use the given θ, the students' response set, and PARAM-1PL to estimate the *b*-value of each item. **Table 4.4** shows the descriptive statistics of the TAKS items' b-

values. Both the TAKS θ and b-values are high (non-negative), indicating that they have been rescaled during test calibration beyond what MLE would have produced from an IRT-1PL model.

The TAKS scales were derived using independent populations: the scales were generated at each grade independently for each domain during test calibration. From now on those scales will be referred to as TAKS referent scales and the θ and *b*-values as TAKS θ and TAKS *b*-values respectively. In theory, the *b*-values across domains and grades should not be compared since the scales are independent of each other. However, since this dataset represents a cohort with longitudinal data and PARAM-1PL allows for simultaneous estimation of both θ and *b*-values by treating the cohort as a test calibration sample, it is possible to compare across grades and domains using the values PARAM-1PL generates natively from the response set. The scales generated are internally referent to the cohort of students under analysis and will be called the internally referent scales and the θ and *b*-values as the internal θ and internal *b*-values respectively. **Tables 4.5** and **4.6** show the descriptive statistics for the PARAM-1PL generated internal θ and internal *b*-values for each domain and year independently estimated using the Analytic dataset. Now that the TAKS and internal values for both θ and b-values have been determined, what is the relationship between the TAKS and internal scales? It has already been shown that with the exception of the Reading domain, the internal θ values are perfectly correlated to the TAKS θ values (see **Table 4.1**). It would be safe to assume then that without the rescaling in Reading, the internal θ values would also be

perfectly correlated to the TAKS θ values in Reading for the multiple choice items only. **Tables 4.7** and **4.8** are the correlation matrix of the TAKS θ and internal θ values respectively. Note how only the Reading domain correlations change. **Table 4.9** shows how the TAKS and internal b-values are also perfectly correlated. This means that it is possible to linearly interconvert between the TAKS and internally referent scales, and that the cohort under analysis and the test calibration sample must be very similar and representative of the testing population in general. This lends credibility to results and conclusions in this dissertation.

The TAKS scales were “anchored” to higher values during test calibration then those generated natively by MLE. The reason for this is unknown, but a possible explanation might be that TEA and PEM wanted all values to be positive so as to not connate negative item difficulty and student achievement (Kline 1993). Regardless, all future analysis will use internal θ and b-values only since PARAM-1PL has issues with the inflated values of the TAKS scales. It is possible that TEA’s software, which is probably not freeware, can handle the TAKS scale, unlike PARAM-1PL. The fact that the scales are perfectly inter-convertible renders issues of which scale to use moot.

The internal scales allow us to make certain comparisons since the Longitudinal dataset represents a case of repeated measures for the same cohort of students across domains. Based on the mean internal b-values found in **Table 4.6** and an analysis of variance, the Reading domain was the easiest, and then followed by History. Math and Science showed no significant difference and

were the most difficult domains on the TAKS exam[‡]. Note that just because the students felt as if one domain was easier than another did not mean that more students were passing the former than the latter domain. The cutoff θ value for passing each section and year was done independently during the field testing of the TAKS exam as explained in Chapter 12 of TEA's *Technical Digest 2005-2006* based on various standards of achievement. So even if the students felt as if a domain was easier, if the cutoff was set higher, fewer students would pass than if the domain was harder but the cutoff was set lower. Critics may argue that one cannot compare across domains since they are each distinct with their own requisite skills and knowledge and so should be orthogonal to each other and in keeping with the assumptions of IRT-1PL that items are only testing for a unidimensional latent trait. The comparison that is being made here is only about the relative ease the cohort of students experienced in each domain based on their response set independently of the content.

The correlation values for the internal θ values as seen in **Table 4.8** are extremely high. This is to be expected intra-domain since it seems likely that a student who is proficient in Math would stay proficient year after year. Of concern, though, are the inter-domain correlations being so high and similar to the intra-domain correlations. Based on the values, students who performed well in one domain would perform similarly across domain for all years. This is contrary to our understanding that the domains each have unique knowledge and skills relative to each other and so should have some level of orthogonality as

[‡] For ANOVA results, please see Appendix C.

opposed to being homogeneous. Critics would explain the shared inter-domain correlation values by saying that students of high proficiency exhibit strong academic motivation in all domains while students of low proficiency exhibit poor academic motivation in all domains. It is this academic motivation that causes students to perform similarly across domains. This explanation is facetious at best. The point of testing different domains is to measure the achievement in each domain, not to measure the overall academic motivation of the students. Recall that the law defines achievement as a measure of how much a student has learned and is able to apply the skills and knowledge defined for each grade level. The goal of accountability is to identify teachers whose pedagogical practices exemplify quality instruction base on student achievement, not academic motivation.

Grade	Domain	T1	T2
9	**Reading**	123.2185	1944.237
	Mathematics	184.6154	2009.908
10	**Reading**	97.06539	1983.745
	Mathematics	141.0437	2038.646
	Science	160.4278	1996.845
	History	145.2081	2046.854
11	**Reading**	113.4816	2017.624
	Mathematics	140.5811	2064.714
	Science	129.4778	2070.868
	History	126.4756	2093.297

Table 4.2 Values for T1 and T2 constants (reproduced from TEA's *Technical Digest 2005-2006*, p. 131)

Domain	Minimum	Maximum	Mean	Std. Deviation
R04	-3.437	9.371	3.345	1.030
R05	-3.624	9.234	3.377	1.126
R06	-2.094	10.005	3.457	1.129
M04	-3.475	5.211	1.020	1.190
M05	-3.436	5.256	0.985	1.196
M06	-3.132	5.380	1.247	1.189
H05	-3.182	5.159	1.809	1.219
H06	-3.427	5.314	1.867	1.234
S05	-3.524	5.293	0.972	1.001
S06	-3.748	5.346	1.003	1.029

Table 4.3 Student TAKS θ descriptive statistics determined using the scales scores from TEA on the Analytic dataset (N = 69,570)

Domain	N	Minimum b	Maximum b	Mean b	Std. Deviation b
R04	33	0.016	3.591	1.131	0.614
R05	48	-0.733	3.235	0.589	0.626
R06	48	-0.599	1.663	0.522	0.500
M04	52	-1.218	1.619	0.219	0.637
M05	56	-1.222	1.284	0.174	0.660
M06	60	-0.908	1.464	0.274	0.641
H05	50	-0.909	1.600	0.437	0.578
H06	55	-1.063	1.597	0.442	0.685
S05	55	-1.257	1.380	0.270	0.608
S06	55	-0.835	1.664	0.300	0.612

Table 4.4 TAKS *b*-value descriptive statistics for multiple choice items of TAKS domains between 2004 and 2006 estimated using TAKS θ from the Analytic dataset.

Domain	Minimum	Maximum	Mean	Std. Deviation
R04	-3.645	3.000	0.165	0.703
R05	-3.703	3.000	0.200	0.698
R06	-3.500	3.000	0.172	0.854
M04	-1.988	3.000	0.125	0.726
M05	-1.931	3.000	0.122	0.726
M06	-1.879	3.000	0.140	0.719
H05	-3.186	3.000	0.215	0.788
H06	-3.381	3.000	0.213	0.795
S05	-2.019	3.000	0.082	0.599
S06	-3.152	3.000	0.085	0.614

Table 4.5 Student internal θ descriptive statistics estimated using PARAM-1PL from only the response set on the Analytic dataset (N = 69,570)

	N	Minimum b	Maximum b	Mean b	Std. Deviation b
R04	33	-1.799	0.368	-0.857	0.515
R05	48	-3.434	0.219	-1.240	0.555
R06	48	-3.442	-0.512	-1.458	0.427
M04	52	-1.714	0.621	-0.490	0.523
M05	56	-1.673	0.383	-0.503	0.536
M06	60	-1.608	0.337	-0.619	0.525
H05	50	-2.020	0.060	-0.877	0.476
H06	55	-3.236	0.014	-0.930	0.577
S05	55	-1.814	0.437	-0.480	0.516
S06	55	-1.436	0.636	-0.477	0.508

Table 4.6 Internal *b*-value descriptive statistics for multiple choice items of TAKS domains between 2004 and 2006 estimated using PARAM-1PL from only the response set of the Analytic dataset.

	R04 TAKS θ	R05 TAKS θ	R06 TAKS θ	M04 TAKS θ	M05 TAKS θ	M06 TAKS θ	H05 TAKS θ	H06 TAKS θ	S05 TAKS θ	S06 TAKS θ
R04 TAKS θ	1.00									
R05 TAKS θ	**0.58**	1.00								
R06 TAKS θ	**0.59**	**0.60**	1.00							
M04 TAKS θ	0.57	0.50	0.50	1.00						
M05 TAKS θ	0.56	0.52	0.52	**0.82**	1.00					
M06 TAKS θ	0.52	0.48	0.51	**0.77**	**0.81**	1.00				
H05 TAKS θ	0.62	0.54	0.54	0.64	0.68	0.61	1.00			
H06 TAKS θ	0.56	0.47	0.49	0.58	0.60	0.60	**0.74**	1.00		
S05 TAKS θ	0.59	0.53	0.52	0.72	0.76	0.70	0.75	0.69	1.00	
S06 TAKS θ	0.55	0.48	0.50	0.68	0.71	0.73	0.70	0.73	**0.77**	1.00

Table 4.7 Correlation matrix for TAKS θ (N = 61,311)

	R04 Int. θ	R05 Int. θ	R06 Int. θ	M04 Int. θ	M05 Int. θ	M06 Int. θ	H05 Int. θ	H06 Int. θ	S05 Int. θ	S06 Int. θ
R04 Int. θ	1.00									
R05 Int. θ	**0.62**	1.00								
R06 Int. θ	**0.59**	**0.62**	1.00							
M04 Int. θ	0.55	0.56	0.54	1.00						
M05 Int. θ	0.54	0.57	0.55	**0.82**	1.00					
M06 Int. θ	0.50	0.52	0.54	**0.77**	**0.81**	1.00				
H05 Int. θ	0.61	0.63	0.58	0.64	0.68	0.61	1.00			
H06 Int. θ	0.56	0.57	0.54	0.58	0.60	0.60	**0.74**	1.00		
S05 Int. θ	0.57	0.60	0.56	0.72	0.76	0.70	0.75	0.69	1.00	
S06 Int. θ	0.54	0.57	0.55	0.68	0.71	0.73	0.70	0.74	**0.77**	1.00

Table 4.8 Correlation matrix for internal θ (N = 61,311)

	R04 TAKS b	R04 Int. b	R05 TAKS b	R05 Int. b	R06 TAKS b	R06 Int. b	M04 TAKS b	M04 Int. b	M05 TAKS b	M05 Int. b	M06 TAKS b	M06 Int. b	H05 TAKS b	H05 Int. b	H06 TAKS b	H06 Int. b	S05 TAKS b	S05 Int. b	S06 TAKS b	S06 Int. b
R04 TAKS b	1.00																			
R04 Int. b	**1.00**	1.00																		
R05 TAKS b	0.04	0.04	1.00																	
R05 Int. b	0.04	0.04	**1.00**	1.00																
R06 TAKS b	-0.03	-0.03	-0.19	-0.19	1.00															
R06 Int. b	-0.03	-0.03	-0.19	-0.18	**1.00**	1.00														
M04 TAKS b	-0.09	-0.09	0.09	0.09	0.02	0.02	1.00													
M04 Int. b	-0.09	-0.09	0.09	0.09	0.02	0.02	**1.00**	1.00												
M05 TAKS b	0.01	0.01	0.15	0.15	0.03	0.03	0.31	0.32	1.00											
M05 Int. b	0.01	0.01	0.15	0.15	0.03	0.03	0.32	0.32	**1.00**	1.00										
M06 TAKS b	0.02	0.02	-0.12	-0.12	0.22	0.22	0.23	0.23	0.47	0.47	1.00									
M06 Int. b	0.02	0.02	-0.11	-0.12	0.22	0.21	0.23	0.23	0.47	0.47	**1.00**	1.00								
H05 TAKS b	0.04	0.04	0.12	0.12	-0.05	-0.04	0.59	0.60	0.49	0.49	0.46	0.47	1.00							
H05 Int. b	0.04	0.04	0.12	0.12	-0.04	-0.04	0.60	0.60	0.49	0.49	0.47	0.47	**1.00**	1.00						
H06 TAKS b	-0.01	-0.01	0.20	0.20	-0.08	-0.09	0.66	0.67	0.46	0.46	0.43	0.43	0.56	0.57	1.00					
H06 Int. b	-0.02	-0.02	0.20	0.19	-0.08	-0.09	0.66	0.67	0.46	0.46	0.44	0.44	0.56	0.57	**1.00**	1.00				
S05 TAKS b	-0.09	-0.09	0.18	0.18	0.10	0.10	0.71	0.72	0.37	0.37	0.36	0.36	0.63	0.63	0.74	0.74	1.00			
S05 Int. b	-0.09	-0.09	0.18	0.18	0.10	0.10	0.71	0.72	0.38	0.38	0.36	0.36	0.63	0.63	0.74	0.75	**1.00**	1.00		
S06 TAKS b	-0.13	-0.14	-0.08	-0.08	-0.17	-0.17	0.51	0.52	0.19	0.19	0.05	0.05	0.32	0.33	0.44	0.44	0.46	0.47	1.00	
S06 Int. b	-0.13	-0.13	-0.08	-0.08	-0.17	-0.17	0.52	0.52	0.19	0.19	0.06	0.06	0.33	0.33	0.44	0.45	0.47	0.47	**1.00**	1.00

Table 4.9 Correlation matrix for TAKS and internal b-values

Student Behavioral Trends

With the internal θ and b-values determined, it was possible to do longitudinal analyses to see how students were behaving across years. As stated before in **Chapter 3**, only Reading and Math were tested for all three years and since Reading experienced a further rescaling based on performance on the essay section, the focus of the longitudinal trends would be on the Math section. The IRT Comparison dataset was used in these analyses since the interest here was only in students who were actually measured by IRT. **Figures 4.1** through **4.4** depict student performance on the TAKS exam for consecutive years. The trend in these graphs was that students tend to perform similarly across years with over 60% of the variance explained by the best fit line. It could be rationalized that students who are proficient in Math generally stay proficient in Math. Note that for all plots used in student behavioral trends, a reduced major axis (RMA) regression was done instead of ordinary least squares (OLS). The reason for using RMA was that since TAKS scores and their related derivations were used as variables, the independent variables were expected to exhibit measurement error. Furthermore, the data was determined to be non-normal as well as heteroscedastic. However, the statistical literature indicates that RMA regression is robust to non-normal and heteroscedastic conditions as long as the variables are linearly related (Warton, 2006). Linearity of the variables was determined by checking for significant higher ordered interactions and none were found.

In order to make more salient how students were changing independent of their scores, **Figures 4.5** and **4.6** depicts student score change between 2005 and 2006 as a function of the change between 2004 and 2005. It can be seen that regardless of their actual performance, students who experienced an increase in their TAKS θ after the first year, generally experienced a decrease in their TAKS θ the following year with the opposite also being true. This lends credibility to the idea that students were regressing to the mean. Critics of the analysis again might protest the fact that the fit line only accounted for 15% of the variance in the data and we reserve argument for the next chapter.

Using the IRT Comparison dataset, **Figure 4.7** shows the Q-Q plot for a normal distribution of the mean change for consecutive years that the students' experienced. Mean change is defined here as the average of the amounts of changes the students experience for each year. Rather than showing the Q-Q plot of each year, the mean was used so there is only one graph. The trend seen does not change when using the mean. The fit i\was not perfect with the tail ends of the distribution "heavier" than a true normal distribution (higher kurtosis). For a true RTM without limits on an interval scale, the expected distribution of changes should be normal if the standard error of measurement is the same for all students (Smith & Smith, 2005). **Figure 4.8** shows the same Q-Q plot except for the fact that the mean change in θ values is now compared to a logistic distribution. Here the fit becomes much better indicating that the most likely situation is that both RTM and ceiling effect were in place. Recall that the floor effect as a possibility has been eliminated logically in **Chapter 3**. Ceiling effect

would cause a heavier accumulation to occur at the tail ends of the distribution since the students in those regions have less room to move about as they approach the limits of the TAKS exam's ability to measure them. To prove this, students in the dataset who scored lower than an internal θ value of -0.5 and higher than 0.5 on any section were removed and then analyzed using the normal Q-Q plot. This is shown in **Figure 4.9**. This range was chosen because it represents the group of students with mean ability that was far from the extremes where the ceiling effect would come into play. Not surprising, the fit to a normal distribution of mean change across years due to RTM is now very good.

While the results of this section are only for Math, the other domains have similar trends even though for the sake of brevity those results were not included. The fact that the students were performing similarly relative to each other every year as indicated by **Figures 4.1** through **4.4**, the fact that students as a population were regressing to the mean independently of their actual scores in **Figure 4.5**, and the fact that the distribution of changes fit a logistic distribution indicate that the students are most likely experiencing RTM relative to their true θ values. This means that the students were not changing their relative rank order to each other or that if they do\id change, the change was monotonic so as to preserve the rank ordering. This is problematic from an educational standpoint since the goal of education is to empower students by teaching them the skills and knowledge that have been deemed important to their functioning in society. The law assumes that achievement is variable and dependent on instructional quality. However, when examining the students across teachers and schools,

student “achievement” does not seem to be variable at all. The next section looks at the items to see if the scales generated for each domain and year were truly independent even though they were generated independently. We already know from the inter-domain correlations that there is definitely no complete independency in the domains.

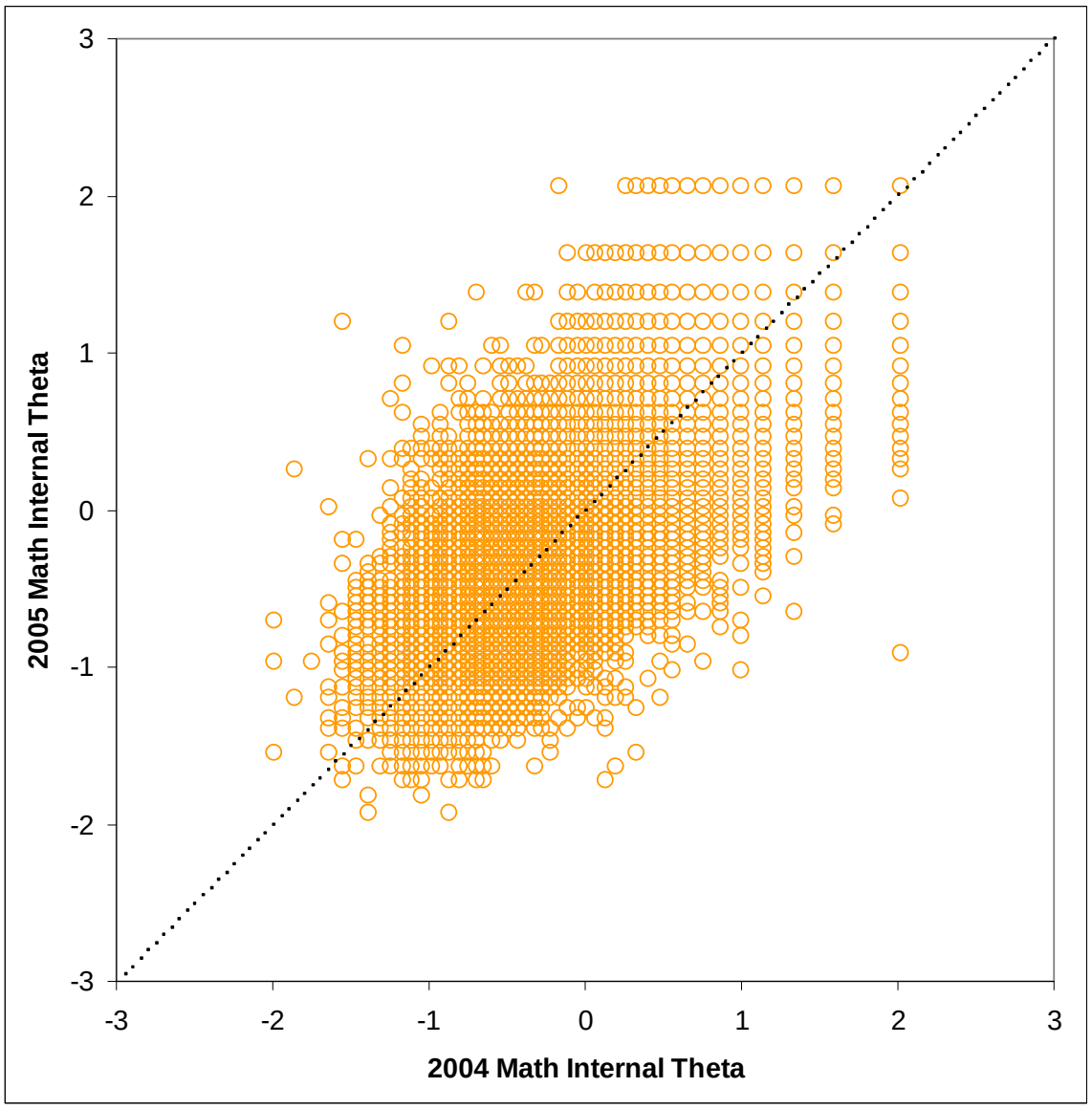

Figure 4.1 Graph of the 2005 Math Internal θ as a function of the 2004 Math Internal θ
(Fit Line: $y = 1.002x - 0.002$, $R^2 = 0.669$)

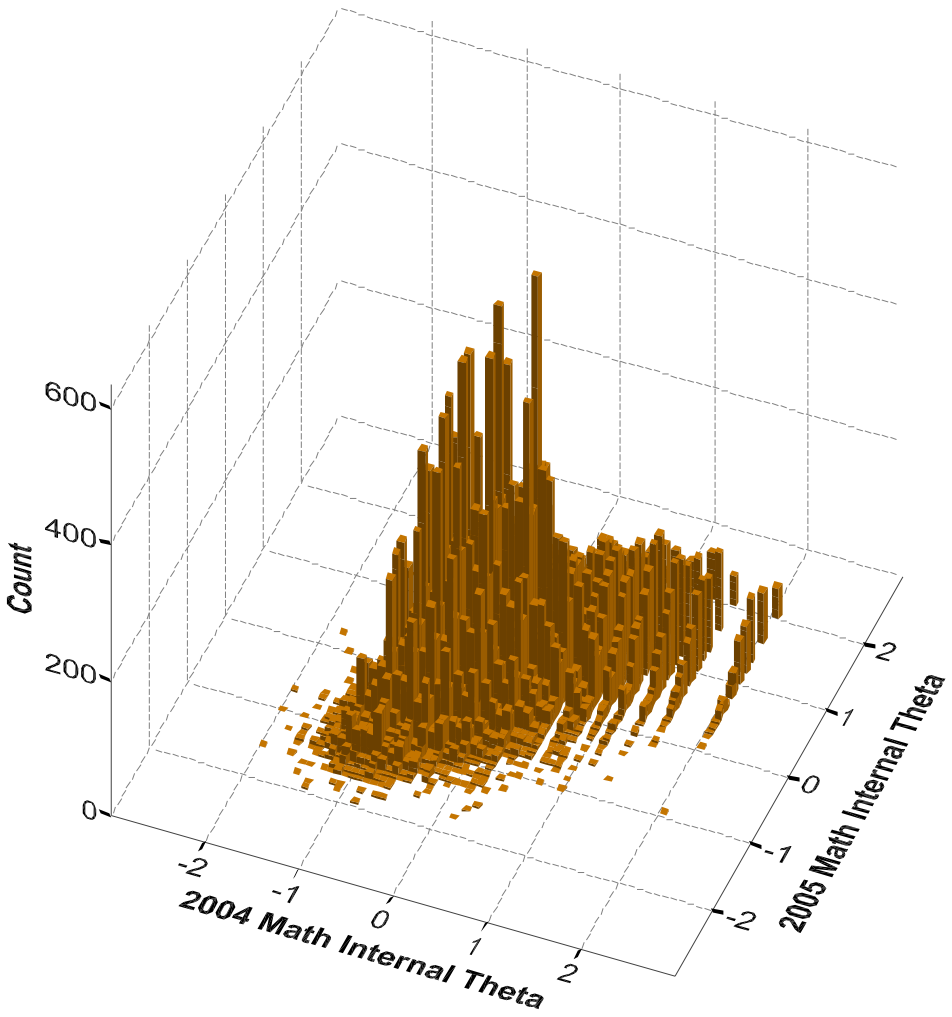

Figure 4.2 3D histogram of the 2005 Math Internal θ as a function of the 2004 Math Internal θ

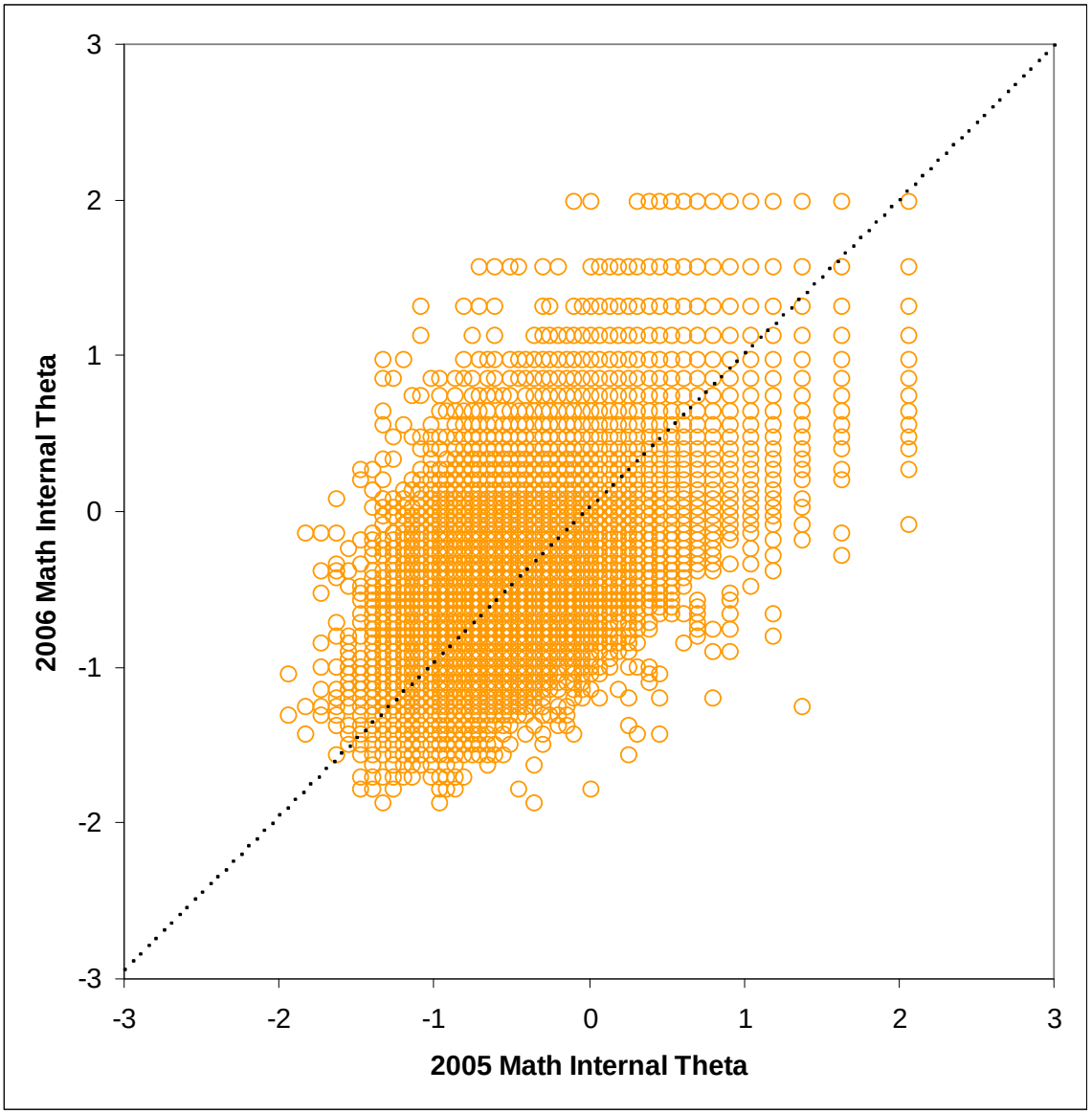

Figure 4.3 Graph of the 2006 Math Internal θ as a function of the 2005 Math Internal θ
(Fit Line: $y = 0.988x + 0.022$, $R^2 = 0.652$)

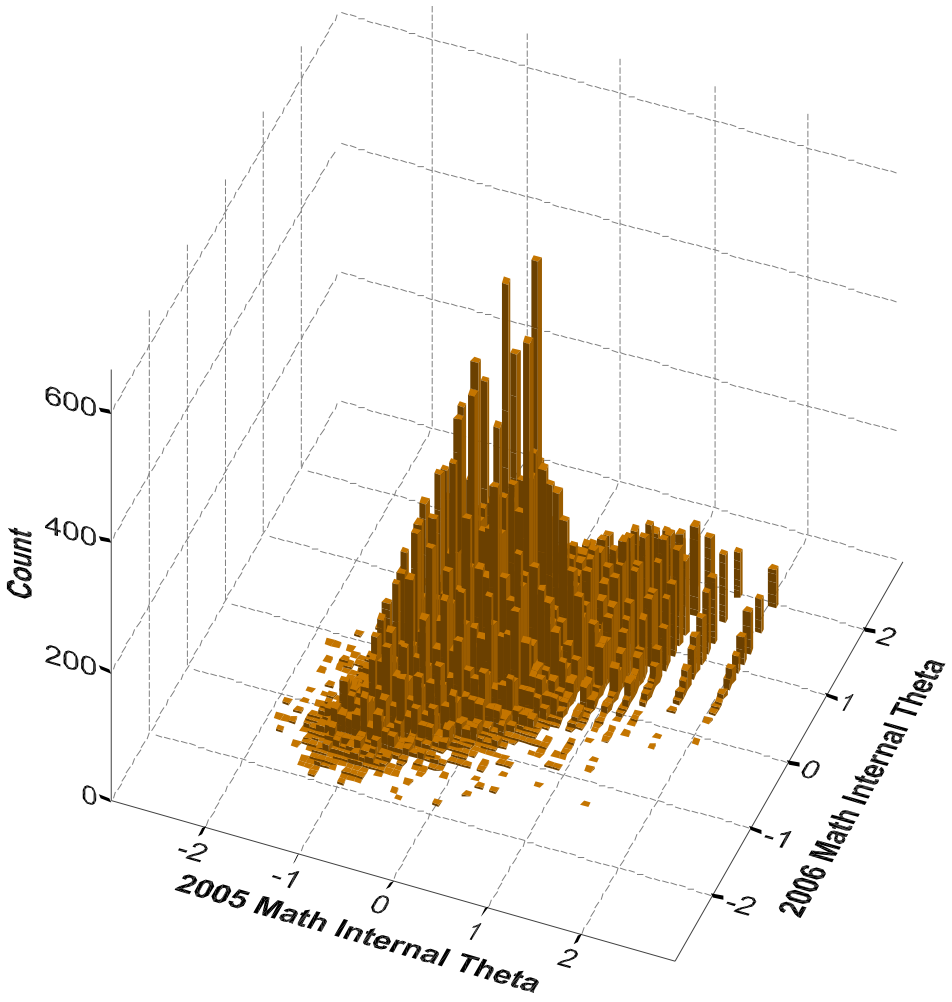

Figure 4.4 3D histogram of the 2006 Math Internal θ as a function of the 2005 Math Internal θ

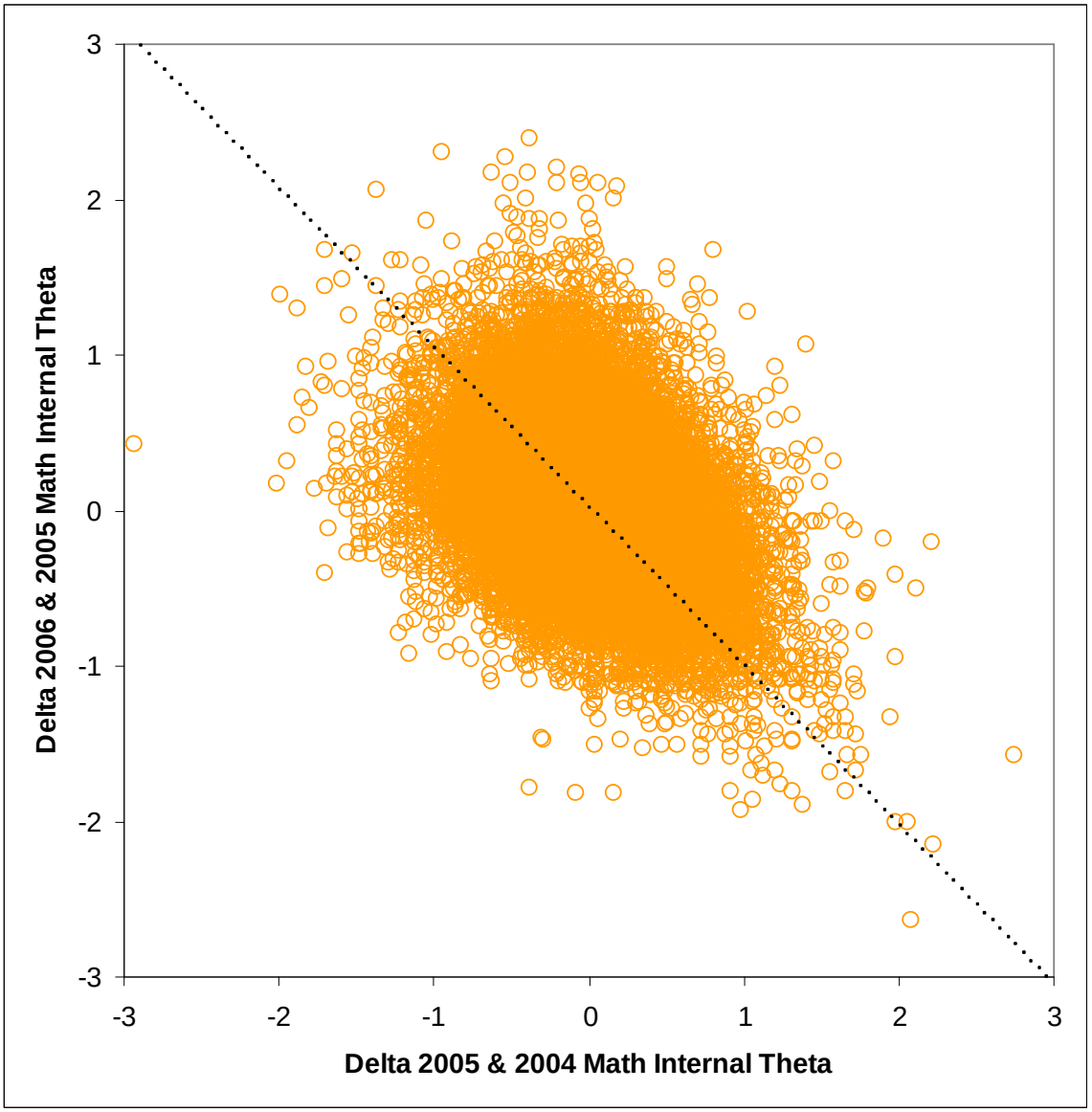

Figure 4.5 Graph of the difference between 2006 and 2005 Math Internal θ as a function of the difference between 2005 and 2004 Math Internal θ
(Fit Line: $y = -1.023x + 0.019$, $R^2 = 0.150$)

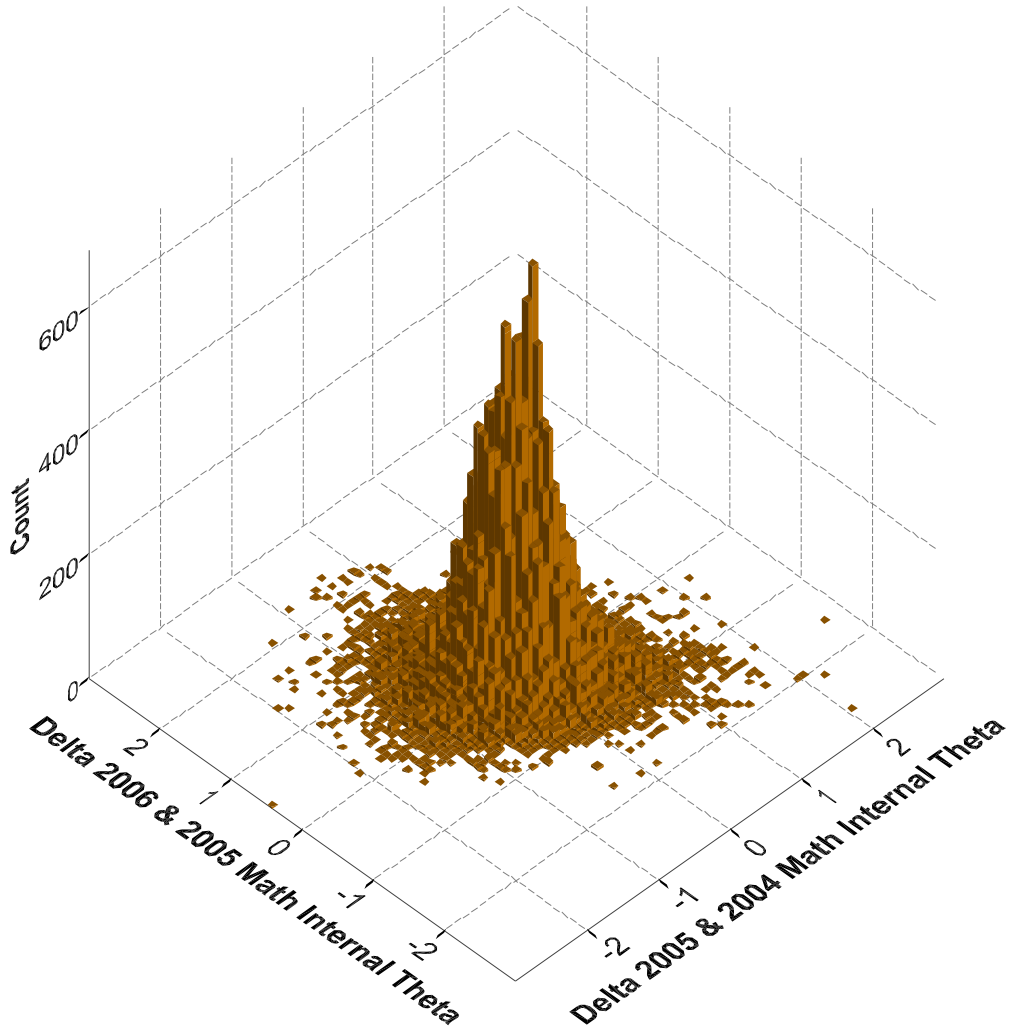

Figure 4.6 3D histogram of the difference between 2006 and 2005 Math Internal θ as a function of the difference between 2005 and 2004 Math Internal θ

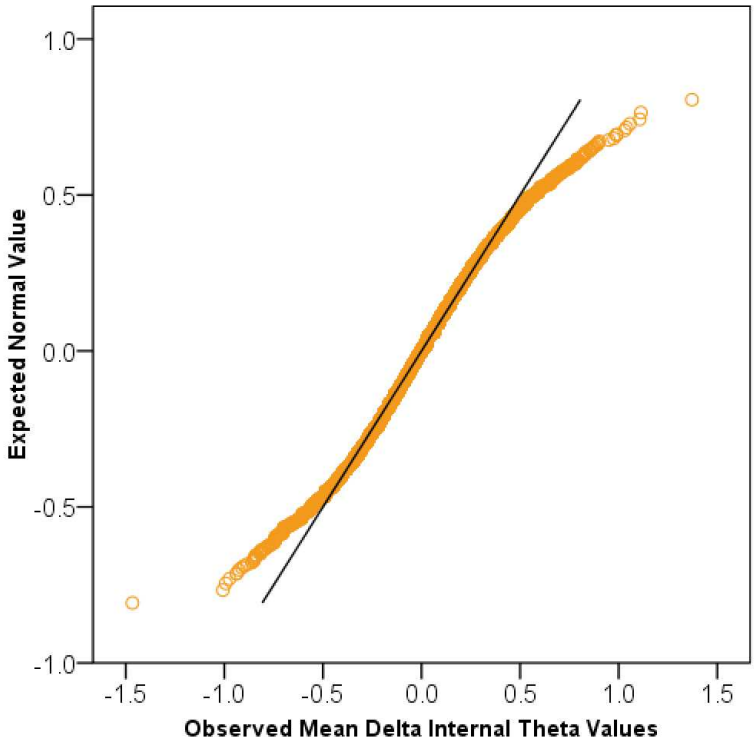

Figure 4.7 Normal Q-Q plot of the mean of the delta internal θ values

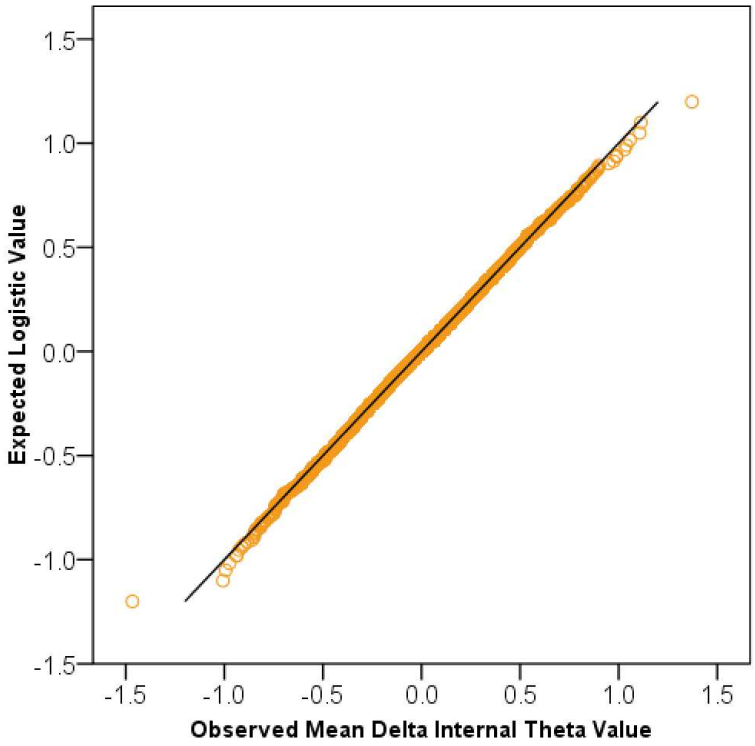

Figure 4.8 Logistic Q-Q plot of the mean of the delta internal θ values

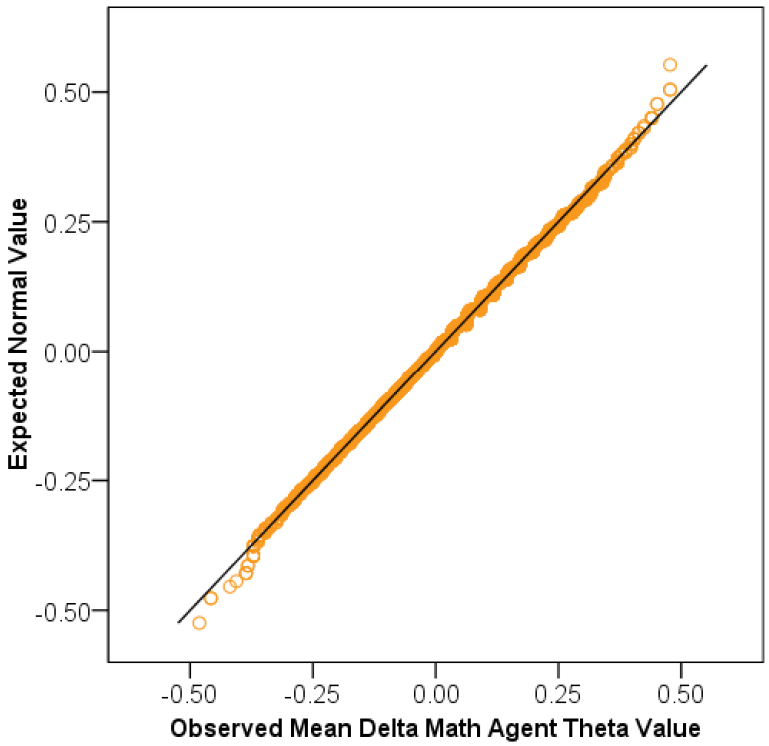

Figure 4.9 Normal Q-Q plot of the mean of the delta internal θ values for students who scored between -0.5 and 0.5 on their internal θ value

Item Behavioral Trends

Table 4.8 shows a high level of correlation both within and across domains for all years, which indicated that students appeared to behave in a similar fashion regardless of domain and year. In order to see just how similar the students were behaving, item response functions (IRF) were generated. To make the similarity even more salient, it was decided that items in each domain of the most recent year, 2006, with an internally referent *b*-value of about -0.90 were used. Items of similar difficulty would have similar IRFs if they are on the same scale. **Table 4.10** shows the items chosen with their internal and TAKS referent *b*-values. **Figures 4.10** through **4.19** show the comparison of the IRFs of the chosen items as a function of each internal scale. As can be seen, within each of the internal scales, different domain items exhibited slightly different IRFs, but overall the different domain items behaved similarly across scales. To further emphasize the point, **Figures 4.20** through **4.23** shows the IRFs of each domain item as a function of all the internal scales. It can clearly be seen that regardless of the scale used, each domain item behaved similarly on all scales.

These IRFs and the correlation matrix of **Table 4.8**, indicated then that whatever the students were being tested for, it was relatively the same across years and domains. We will call this unknown entity the Profiling mental latent trait (LTP) since it cannot be domain related. By not giving it a more specific name than LTP, it was hoped that any bias will be mitigated. Profiling was chosen because we have shown that students were being persistently rank

ordered during the Longitudinal Trends analysis. Here, we have shown that the different scales of θ values a\were actually not so different and seem to be measuring a common construct. Whatever LTP may be, it was causing a profile to arise in the students regardless of year or domain. LTP may be related to racial, cultural, socioeconomic status, or any other number of ways in which we can disaggregate students. We leave that discussion for other researchers.

Note that in each IRF figure, the y-intercept is slightly above 0.80. This is due to the fact that all items have an internally referent *b*-value of approximately -0.90. If a higher *b*-value was chosen, the y-intercept would be smaller and if a lower b-value was chosen, the y-intercept would be higher. The *b*-value of -0.90 was chosen simply because it was closest to the mean *b*-value of all the different sections. The lower extremity of all the IRFs have some level of stochasticity since there were very few individuals at those θ values and if one of them responded correctly on the item, it would cause the ratio of correct response to be higher than probability would dictate on the IRF. Theoretically there should also be very few individuals at the upper extremity as well. However, due to the easy nature of the TAKS exam, a larger number of students have scores in the upper range and hence the lack of stochasticity there. Of all the scales, the Reading ones exhibited the most stochasticity since Reading was the easiest domain based on the mean internal b-values for this sample.

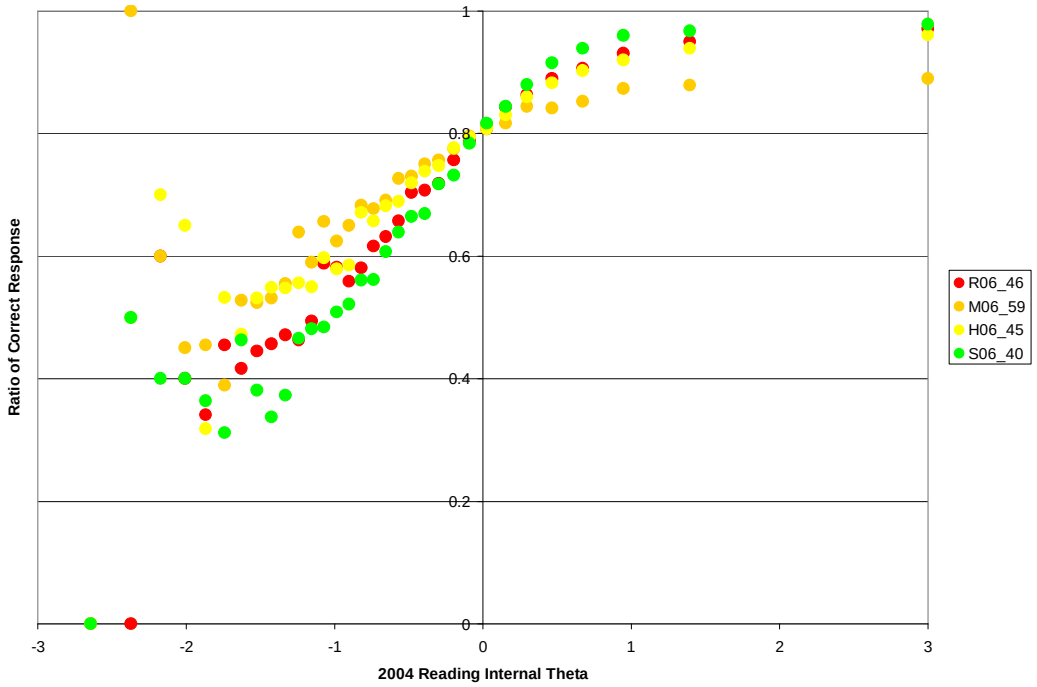

Figure 4.10 IRFs of the varying domain items to the 2004 Reading scale

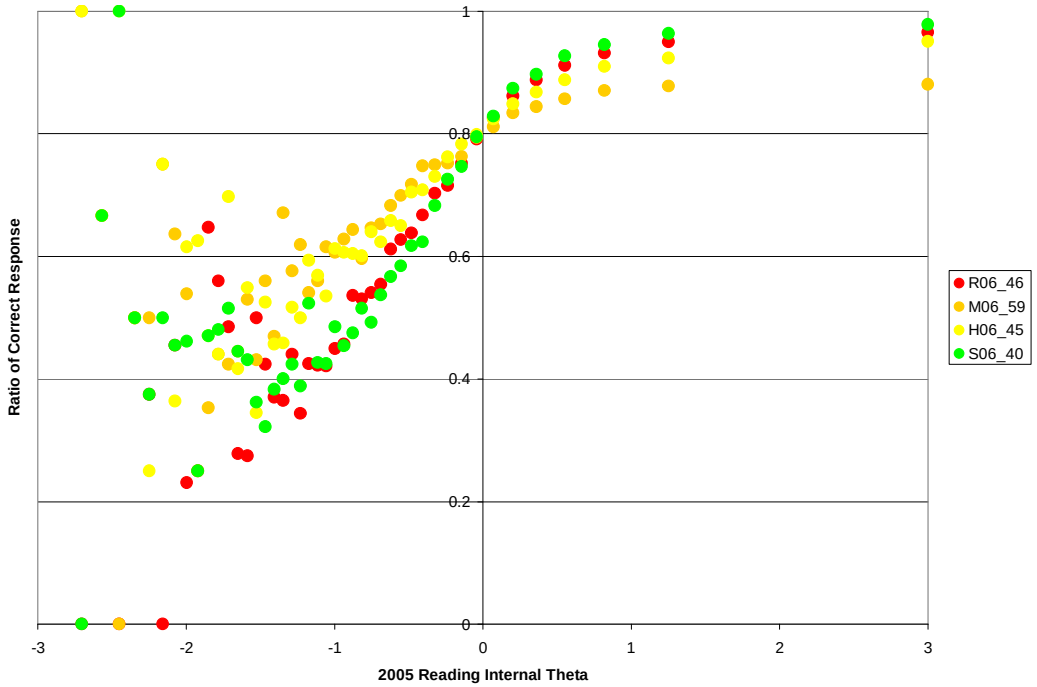

Figure 4.11 IRFs of the varying domain items to the 2005 Reading scale

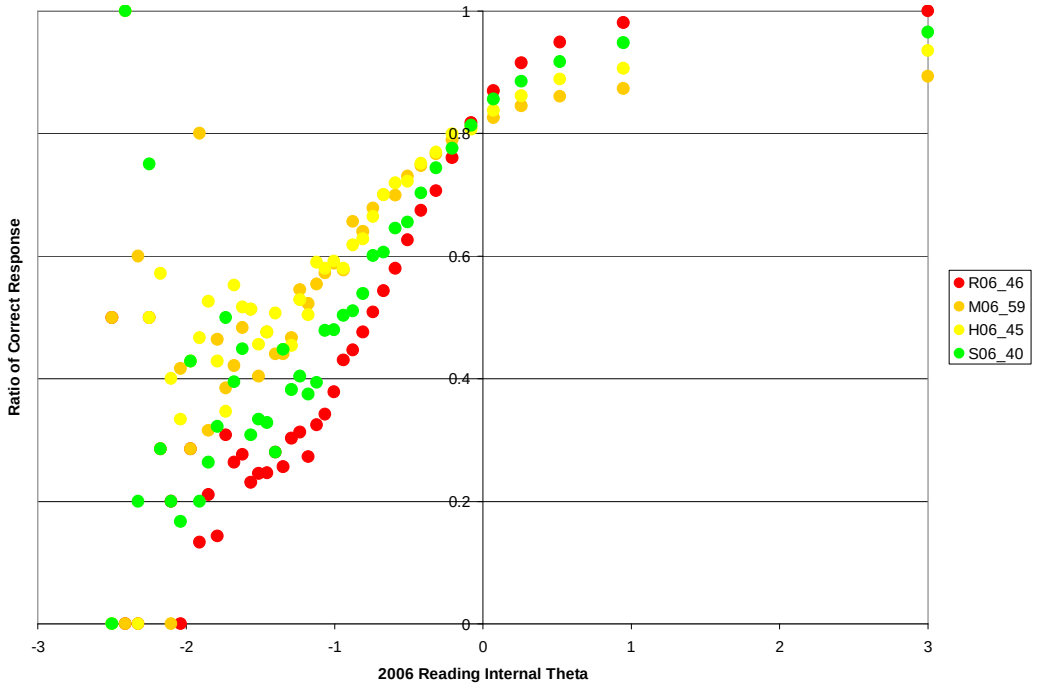

Figure 4.12 IRFs of the varying domain items to the 2006 Reading scale

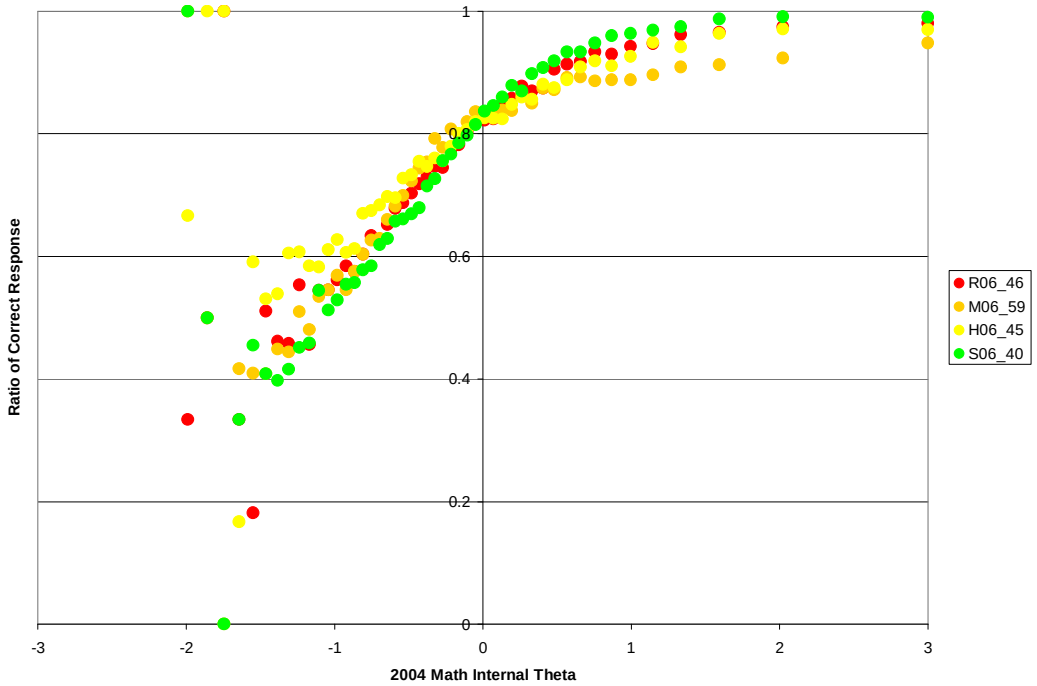

Figure 4.13 IRFs of the varying domain items to the 2004 Math scale

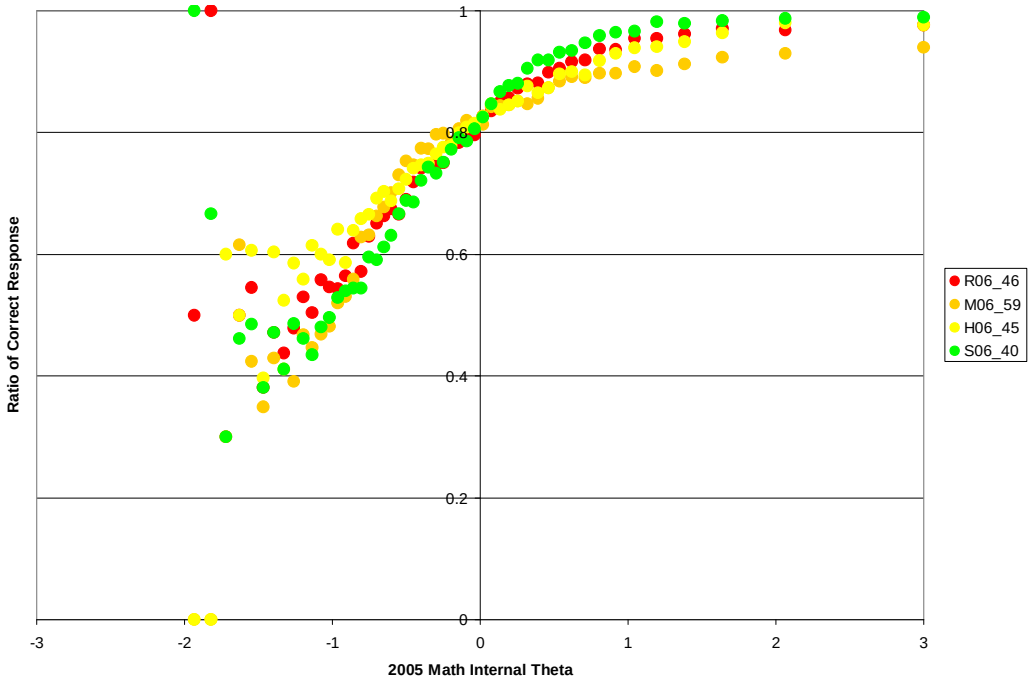

Figure 4.14 IRFs of the varying domain items to the 2005 Math scale

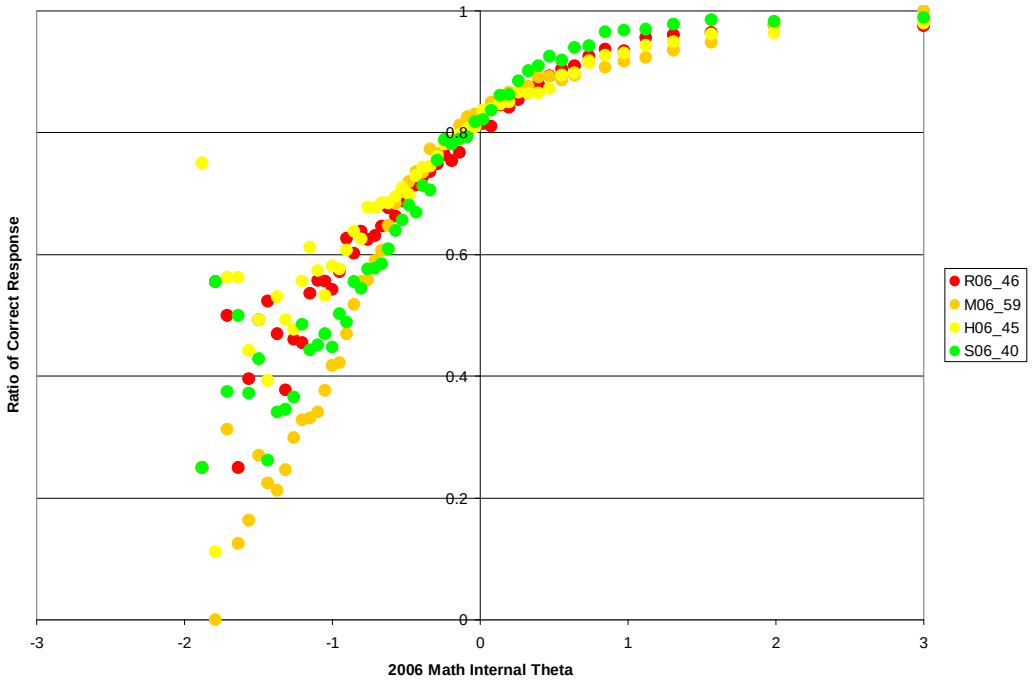

Figure 4.15 IRFs of the varying domain items to the 2006 Math scale

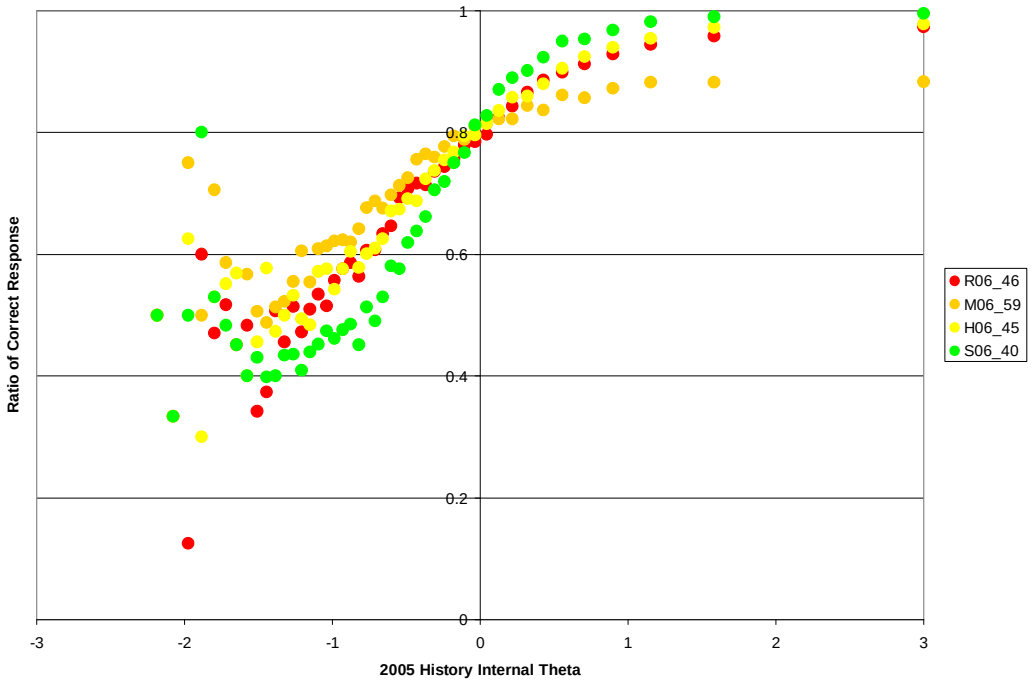

Figure 4.16 IRFs of the varying domain items to the 2005 History scale

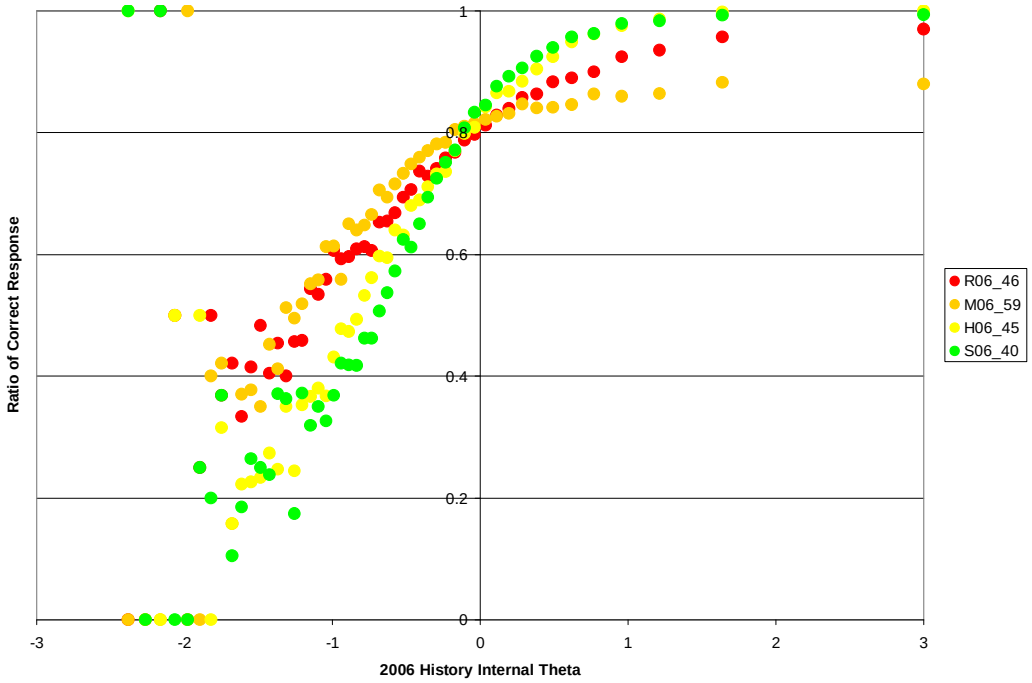

Figure 4.17 IRFs of the varying domain items to the 2006 History scale

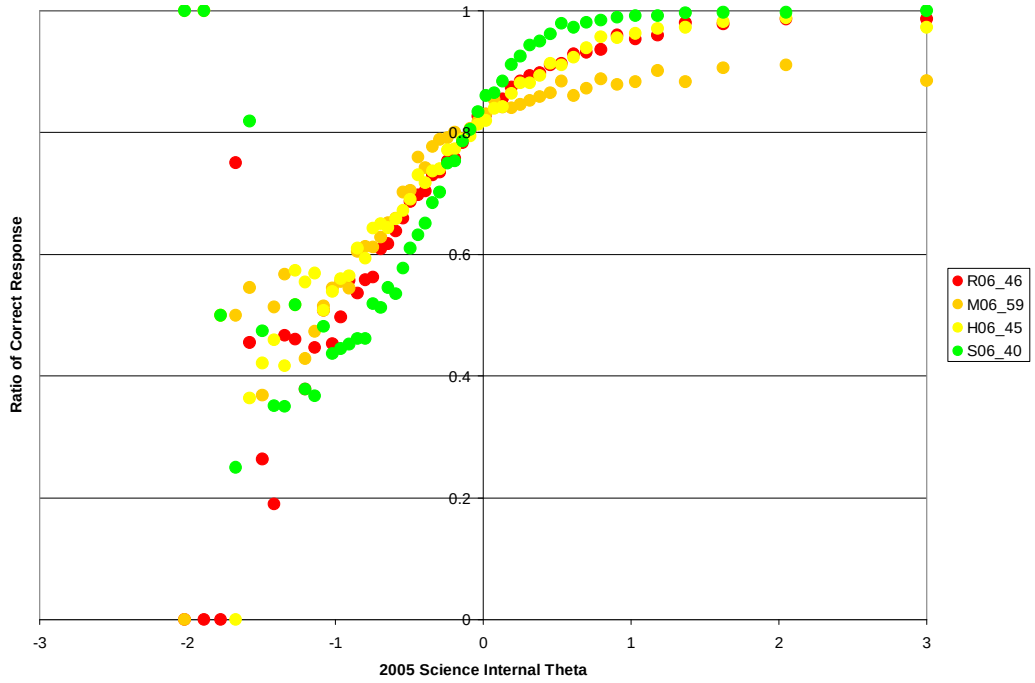

Figure 4.18 IRFs of the varying domain items to the 2005 Science scale

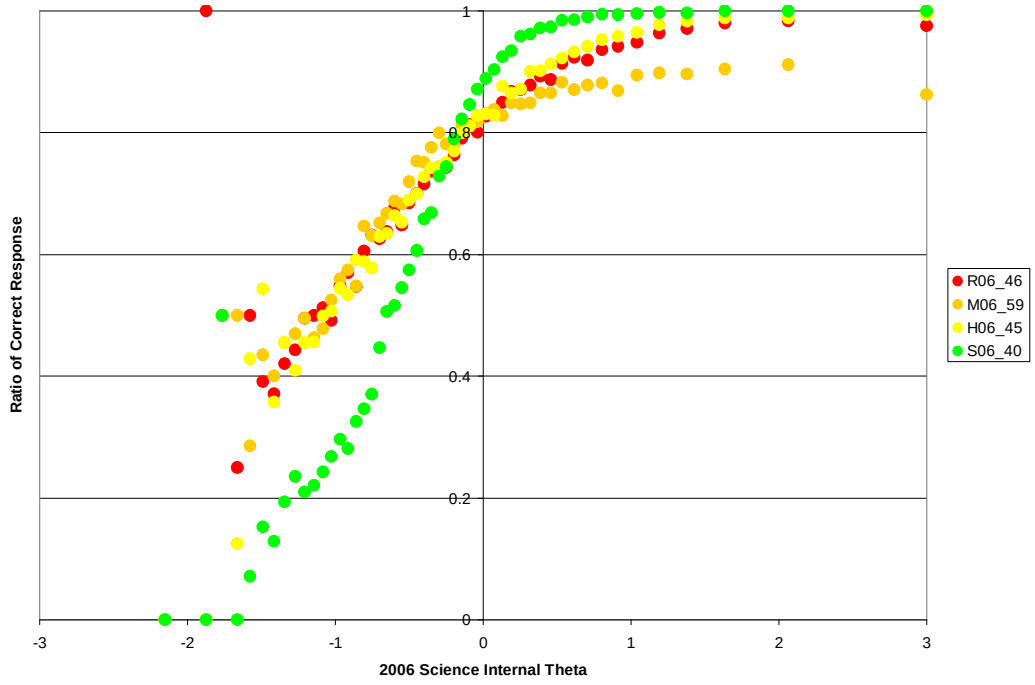

Figure 4.19 IRFs of the varying domain items to the 2006 Science scale

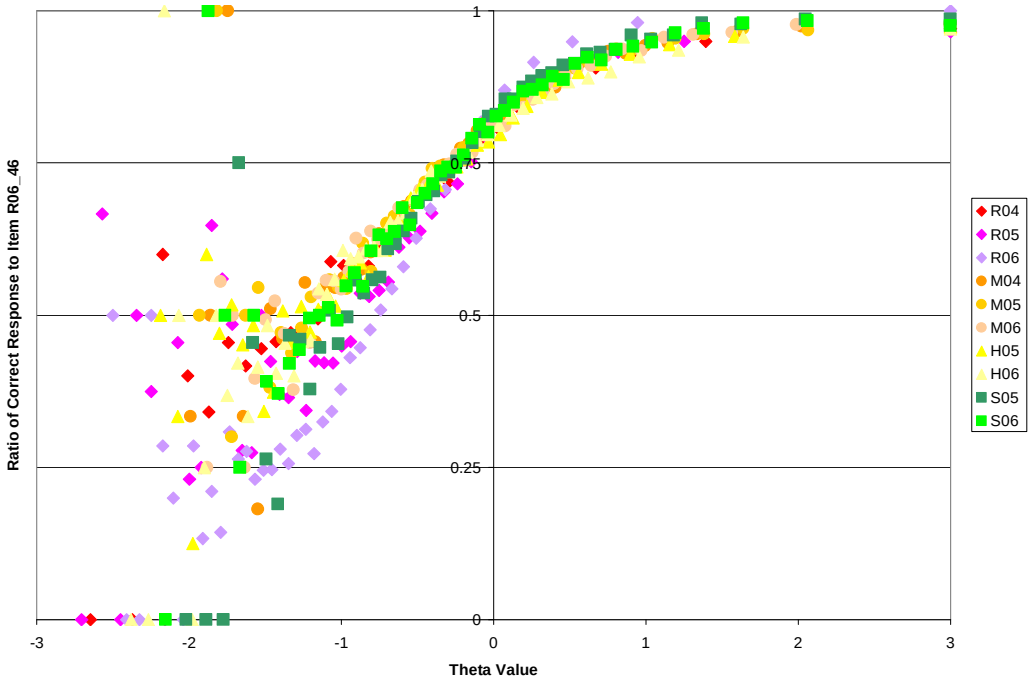

Figure 4.20 IRFs of item R06_46 as a function of the different internal scales

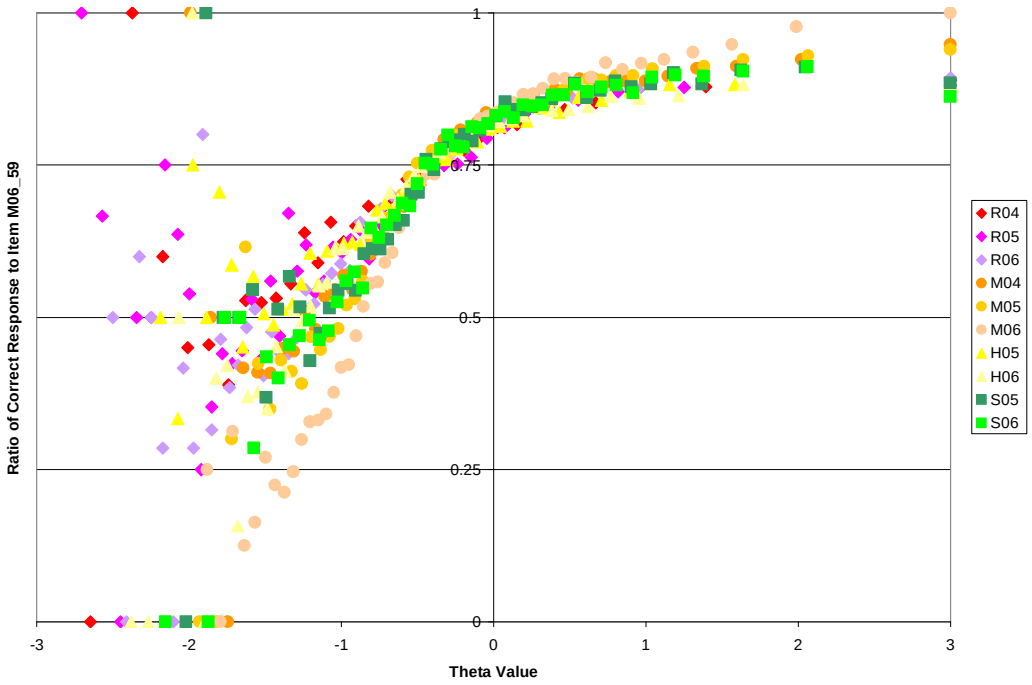

Figure 4.21 IRFs of item M06_59 as a function of the different internal scales

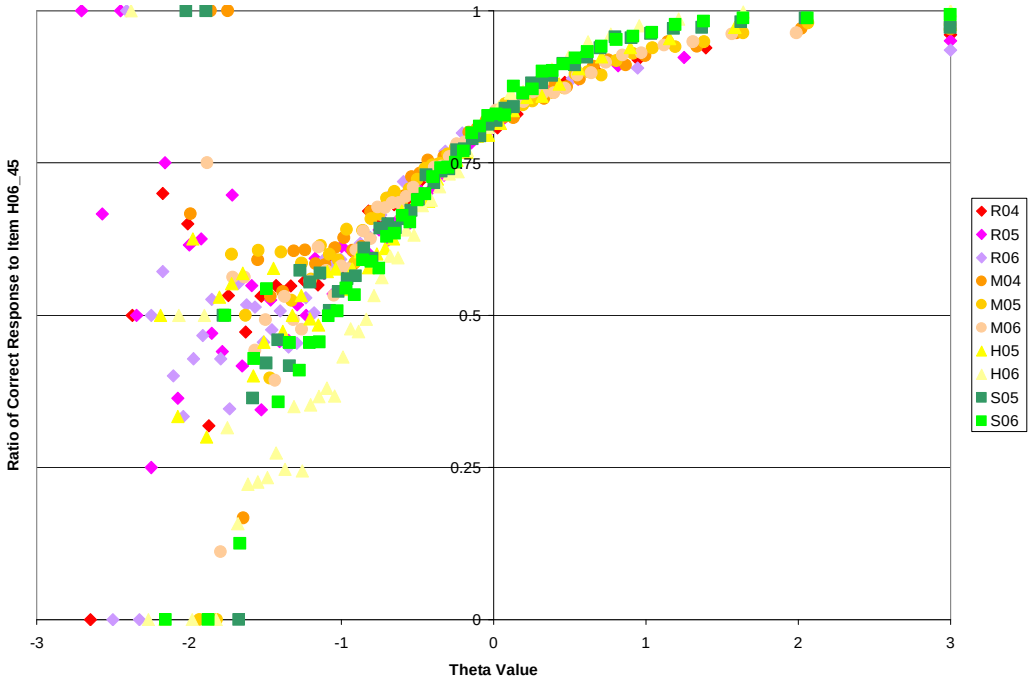

Figure 4.22 IRFs of item H06_45 as a function of the different internal scales

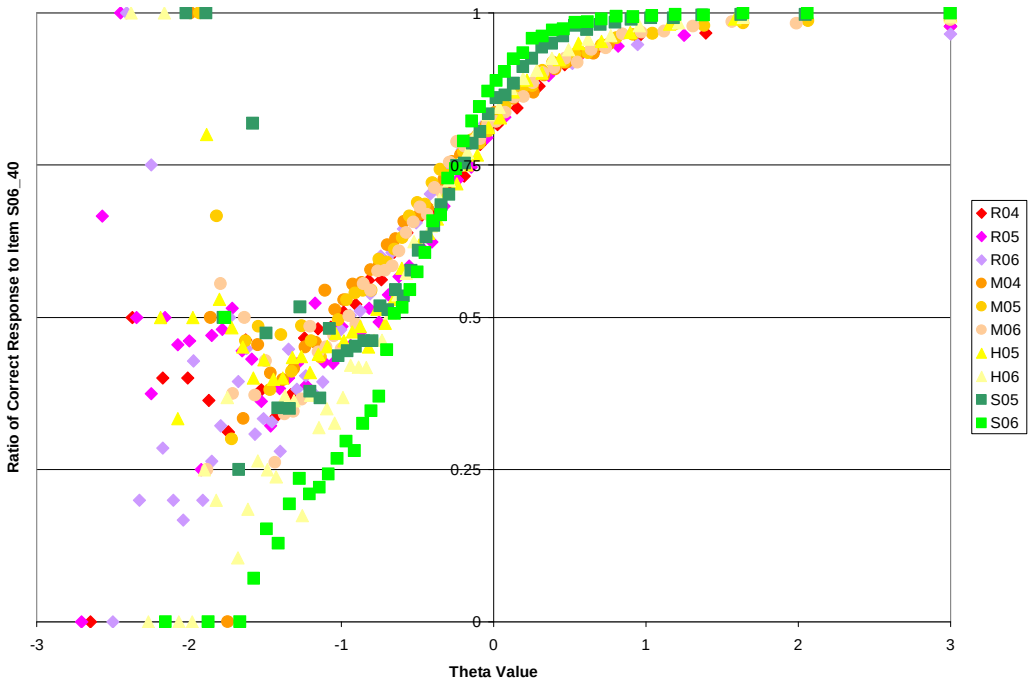

Figure 4.23 IRFs of item S06_40 as a function of the different internal scales

Domain	Item #	Internal b	TAKS b
R06	46	-0.919	1.162
M06	59	-0.889	-0.062
H06	45	-0.907	0.454
S06	40	-0.929	-0.249

Table 4.10 Items chosen for item analysis

Student and TAKS Exam Interaction

It has been shown that the different domains share a large proportion of the variance among themselves such that the IRFs were pretty much the same regardless of the scale. This shared variance across the domains represented a common entity that has been called the Profiling mental latent trait (LTP). In order to know how much of an influence this latent trait plays, a structural equation model (SEM) analysis was done. SEM required that a model be constructed *a priori* to the analysis and should be theoretically informed.

Figure 4.24 shows the results of the SEM analysis for the IRT Comparison dataset. All variables in boxes were actually measured and all variables in bubbles were estimated from the data. The numbers on the arrows are the correlation values and the numbers in the boxes and bubbles represent amounts of variance. The numbers in bubbles labeled with XRES with X representing one of the four domains are the residual variance that is not explained by LTP. The ER# terms are amounts of variance not accounted for by the domain latent traits on each section. The domains were represented as latent traits in the model. The model assumed that the shared variance on sections between years within a domain represented the common knowledge within that domain being tested. Furthermore, the shared variance between the domains latent traits then must represent the Profiling mental latent trait. Reading, Math, History, and Science are very distinct domains with different knowledge and skills

associated with them. It was logical then to assume that only the unique variance within each domain latent traits represented these unique knowledge and skills.

For the different indices of goodness-of-fit, the χ^2 value was very large (12,870, $p < 0.001$) which would indicate poor fit. Given the sample size though ($N = 61,311$), we should also rely on other measures for goodness of fit. The Root Mean Square Error of Approximation (RMSEA) has a reasonable value of 0.082 while all of the incremental fit measures showed excellent fit values of above 0.90. Overall, the model was accepted as a reasonable one of how the variances were shared between all of the variables in the proposed model.

A large proportion of the variance in each section across years within a domain could be accounted for by the domain latent traits. Furthermore, an astonishing amount of the variance in each domain could be accounted for by LTP. In fact, all of the variance in the Science domain latent trait could be explained by LTP. The amount of variance in each section not explained by the domain latent traits was thought to come from two separate sources: random measurement error and unique knowledge being tested on that section only. Similarly, the residual variance in each domain latent trait not explained by LTP was thought to be comprised of some small random measurement error and the unique knowledge and skills in that domain. It should be mentioned that in theory, there should be very little measurement error in the domain latent traits since they were estimated from shared variances across sections in each domain.

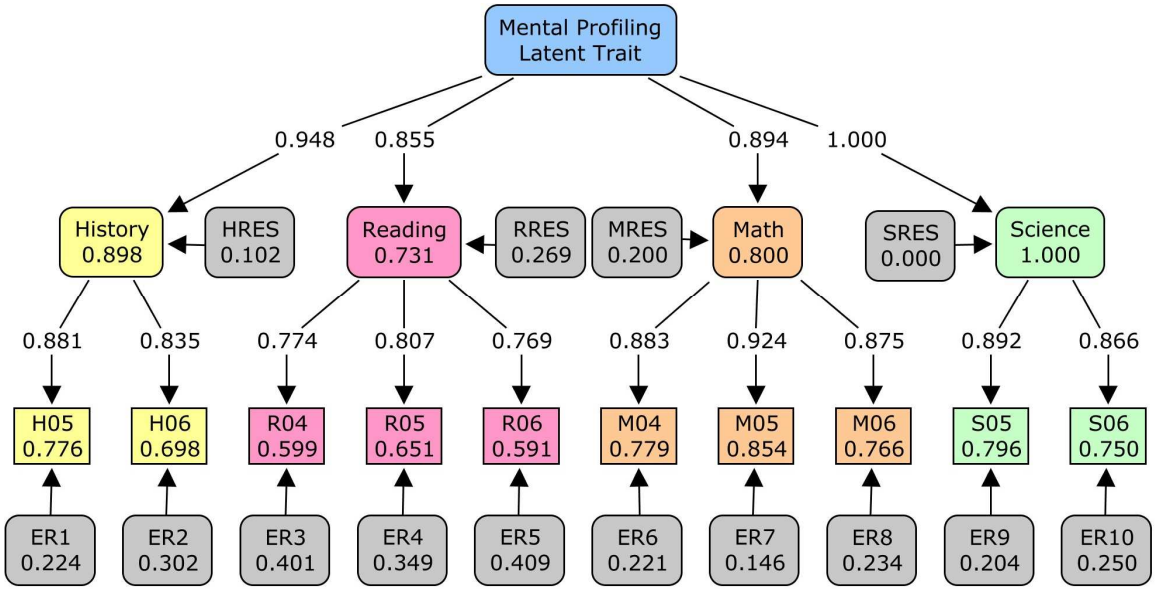

Figure 4.24 Structural Equation Modeling of Internal θ using IRT Comparison dataset (χ^2 =12,870.146 ($p < 0.001$), N= 61,311, df = 31, RMSEA = 0.082, TLI = 0.963, NFI = 0.975, RFI = 0.963, IFI = 0.975, CFI = 0.975)

Discussion

The analysis of the real world data has yielded some interesting results. First and foremost was the fact that using the Analytic dataset as if it were the test calibration sample for the TAKS exam yielded both θ and b-values that were perfectly correlated to values determined by PEM and TEA. One would expect there to be at least some slight deviation from the TAKS scales, no matter how minor. However, within the limits of the four significant digits in the data, the correlations were perfect. There are two conclusions that can be drawn from this fact. First is that the item and total score invariance principle required by IRT has been met on the TAKS exam and second is that the Analytic dataset and the test calibration sample were not statistically different and probably either were a) representative of the testing population in general or b) both the Analytic dataset and the test calibration sample have been processed similarly and thus selected for the same statistical group of students. Even though the raw data had be processed, the processing did not affect the population distribution enough to skew any estimated values from those determined by PEM and TEA and lent credibility to the various analyses in this dissertation.

One of the assumptions of IRT is that each item only measures a unidimensional latent trait for a given θ scale. However, the item analysis indicated that this was not the case at all. The different domain θ scales behaved more or less in the same manner. Thus, TEA and PEM may have labeled the θ scales as being different based on domain face validity, but they did not behave

as such. If we assume that the items must have some level of content validity, the question is whether the items were behaving similarly because they were measuring one common latent trait across domains, or if they were measuring the domain latent trait of interest as well as other latent traits that were shared across the domains. The SEM analysis indicated that each domain did possess some unique variance not explained by LTP. Since the estimation of latent traits in SEM used the shared variance from the individual TAKS sections that were actually measured, the domain latent traits were not as affected by random measurement error. If we accepted that each domain possesses unique skills and knowledge relative to each other, then the residual variance in each domain latent trait that could not be explained by LTP in the SEM would represent those unique skills and knowledge. If that was truly the case, it would mean that the TAKS exam measured for very little domain specific knowledge and skills. Of all the domains, Reading had the most domain specific content, followed by Math and History, respectively. Science had no domain specific content according to the SEM.

The longitudinal trends indicated that students were not changing their scores relative to each other, but rather were maintaining a relative rank ordering such that even if they were growing, they did so in a monotonic fashion. No meaningful change was actually observed in the students over the time period of the analyses of three years. Any changes observed in the scores from year to year are the result of random fluctuations about the mean due to RTM. This might be the result of the invariance requirement of IRT. Recall that the

invariance requirement of IRT means that item b-values and student θ values do not change. Again if we assume that achievement is somehow affected by the varying pedagogical practices of teachers, this would mean that the θ values would have to be insensitive to the differences between those practices. We have all ready quoted that instructional effects were considered a threat to invariance by psychometricians. This has a lot of implications in our current high stakes environment of educational accountability. If students are being persistently rank ordered in a manner that is insensitive to instructional effects then holding students, teachers, and school systems accountable for something out of their control would be detrimental. Furthermore, issues of equity are involved since students were not changing their rank ordering. This would maintain any achievement gap present in the students prior to test calibration.

When the TAKS exam is administered, a great deal of qualitative descriptors on the students was also collected. These qualitative descriptors were self reported by either the parents or the students. While the focus of this dissertation is not on the social inequities of standardized testing, it is important to point out certain trends in the data to emphasize the importance of further research into how standardized testing works. **Figure 4.25** shows the distribution of standardized 2004 Math scale scores within each ethnic group for the Longitudinal dataset (N = 139,062). Note that the percentage values denote only within the group and not as a part of the whole population for each ethnic group. It is interesting to note that the bimodality of the Math score distributions seen during data processing can now be explained by scores disaggregated by

ethnicity. The clustering of Hispanic and Black students at a lower mean than White and Asian student at a higher mean contributes to the bimodality. While the results of the other years in Math and the other domains are not shown, they show no deviation from the established patterns of the 2004 Math section. The only difference was that the mean differences between the ethnicities were smaller in the other domains. **Figure 4.26** shows the relative contribution of each ethnicity to the total distribution of 2004 Math scale scores. In terms of percentage of the student population in the Longitudinal dataset, White students are the most numerous followed by Hispanic, Black, and Asian respectively. **Figure 4.27** shows the achievement trends longitudinally by ethnicity as well as each year's cutoff for "Met Standards" and "Commended" performance. Asians have a mean achievement of almost one standard deviation above the mean of the entire population. White students' mean achievement is about 0.3 standard deviations above norm. Hispanic and Black students hover around -0.4 and -0.5 standard deviations, respectively. We already know that the TAKS exam rank orders students persistently based on some intrinsic latent trait that we have called LTP. It is not surprising then that the achievement gap stays stable across years if the differences were already present in the testing population before test calibration. Rather acting as an agent to minimize the differences between ethnicities as is the stated goal of NCLB, the TAKS exam actually keeps those differences static. Incidentally, it is interesting that the performance cutoffs moved around quite a bit from year to year even though the students were not.

More could be said about this, but it is outside the scope of interest of this dissertation.

Figure 4.28 shows the distribution of standardized 2004 Math scale scores within each socioeconomic status (SES) classification for the Longitudinal dataset. The classification is given in **Table 4.11**. The bimodality of the raw distribution can also be explained as due to the linear combination of the different SES groups having different means. **Figure 4.29** shows the relative contribution of each SES to the total distribution of Math scale scores for 2004. **Figure 4.30** shows the longitudinal achievement trend disaggregated by SES. The general trend is that if students were economically disadvantaged, they tended to perform worse than students who were not regardless of their specific level of being disadvantaged.

Lastly, **Figure 4.31** and **4.32** are similar to the ethnic and SES analyses above except using the gender data. The graph of the relative contributions to each scale score by gender is not shown since it is not as informative with the sample being almost evenly split in half. It can be seen that male students perform better on Math and than female students and that this difference, as expected, was persistent across years. However, unlike ethnicity and SES, the gender analyses changes across domains. For Science and History, male students also performed better. These analyses are not included since they do not differ from the Math ones. However, female students performed better in Reading than male students as shown in **Figures 4.33** and **4.34.** This provides evidence that there is some level of orthogonality in the different domain scales.

Reading here exhibited the most orthogonality compared to the other domains. This could be part of the reason why it had the most amount of unique variance of all the domains in the SEM. These results also validated the assumption made in the computer model that the domain latent trait value determines the scores within each domain for all years.

LTP has thus far been undefined except that it profiled students consistently across domains and years. Here we have discussed three possible components of LTP: ethnicity, SES and gender. There are probably other factors that contribute to LTP that were not collected with the TAKS exam. The fact that the TAKS exam ranked students on LTP made it no better than norm-referenced exams, even though the TAKS was supposed to be criterion-referenced and therefore better. A better label perhaps should be that they are both cutoff referenced with the cutoff being arbitrary from year to year. The only real difference between the norm- and criterion-referenced is that one rank orders students on a normal curve and the other against the original test calibration.

Critics will point out that students at different schools will experience different environments while students at the same school will experience the same environment. School environment is defined here as the set of the unique combination of the different levels of all the factors that students are exposed to when going to school, including instructional quality. School environments tend to be relatively stable over time. It is possible then that the differentials in student TAKS scores are a result of differential school environments while the stability of student TAKS scores across years are a reflection of the stability of the school

environment. It is important then to show that school environment does not affect TAKS scores if we wish show that the TAKS exam is in large instructionally insensitive since instructional quality is a part of the school environment. Note that instructional sensitivity is defined here as responsiveness to the varying pedagogical practices of teachers and allows for standardized testing to be use as an accountability tool.

However, there are complex issues at play. Sex, ethnicity, and SES are factors that are intrinsic to the students and therefore independent of school environment, but the makeup of the student body at a school contributes to that school's unique environment. This makes separation of the effects of student intrinsic factors and schools intrinsic factors difficult. That is why we have used a three pronged approach to examining the TAKS exam. Of the three approaches, item behavioral analysis indicates that regardless of domain label, items behave in a similar fashion. The conclusion that was drawn from this is that the different TAKS domains tested for the same construct that we have called Profile. If the TAKS exam was instructionally sensitive, the items would not behave so similarly across domains since the domains are different, assuming that teachers do teach their designated domain content area.

Figures 4.35 and **4.36** show how SES and ethnic makeup were distributed in 2006 among the different classification of schools in Texas as put forth by TEA based on various factors such as drop out rates and TAKS scores. Note how certain trends correspond with "achievement" well. For instance, the percentage of White students increased with better school rating while the

percentage of Hispanic and Black students increased with decreased school rating. For SES, the percentage of students who were not economically disadvantaged increased with school rating and the percentage of Free and Reduced Lunch students increased with decreased school rating. As student intrinsic factors, SES and ethnicity should be independent of instructional quality, a school intrinsic factor, unless one makes a claim of racial and SES prejudices by the school system. With distinct localized populations occurring at schools of different rating, a statistical analysis should indicate that the school the students attended affects their TAKS scores which can be taken that instructional quality matters on the TAKS exam. This is contrary to the claims of this dissertation that it was actually the profile of the students that was the cause of the differential scores.

Regardless, a four-way ANOVA with ethnicity, sex, SES, and campus rating as factors with the 2006 Math scale score as the dependent variable was done. It was assumed that if the TAKS exam was sensitive to instructional quality, then the campus rating based on TAKS scores would be a measure of the instructional quality at that campus. Campus rating is the rating of the physical school that the student attended. The 2005 rating was used since it provides a historical perspective of how the school fared in the past and limited the effects of circularity since the 2006 TAKS scores would determine the 2006 campus rating. A few things should be noted. The analysis did not use the complete set of students in the Longitudinal dataset. Campuses that had fewer then 200 students were eliminated. This was done to both ensure that sample

size within each group was large and allowed SPSS to be able to process the dataset. This could cause some bias to occur in the data. However since N was about 100,000 students and analysis of demographic distribution indicated that student demographics were not affected by this processing, it was concluded that the processing should have little effect on the results. Also, the sample was not normally distributed, but as many statisticians have noted, for large sample sizes ($N > 100$), this should not be too much of a concern as the nominal and actual α values will be very close. Lastly, the groups exhibit heteroscedascity by Levene's test. To account for possible inflation of actual α value due to heteroscedascity, a more stringent α level of 0.01 was used as suggested in the statistical literature (Stevens, 1999). Ideally, in such a case as this where there is non-normality and heteroscedascity, nonparametric methods should be used such as the Kruskal-Wallis or Friedman's test. This proved ultimately untenable due to the large sample size and number of groups that SPSS would not allow those tests to process.

The ANOVA was first run with a full factorial and then rerun with only significant factors and interactions. The results are displayed in **Table 4.12**. Of the different factors, sex was the easiest to interpret since it has a significant main effect with no interaction effects; that is, male and female students performed differentially, independent of ethnicity, SES, and campus rating. The persistent effect of gender across campuses is indicative of a Profiling latent trait since public schools are coed with both male and female students in the same class and therefore should receive the same quality of instruction. The remaining

three factors all have interaction effects so those need to be examined individually. The ethnicity and campus rating coding system used is in **Table 4.13**. The SES coding is from **Table 4.11**.

Figure 4.37 is the plot for the ethnicity and SES interaction. The trends for SES were different for the different ethnic groups resulting in a significant interaction effect. In particular, the Reduced Lunch and Other Forms of Economically Disadvantaged students exhibited varying results across the ethnicities. The interaction between campus rating and ethnicity was also significant indicating that the different ethnicities behaved differently when at different campuses as shown in **Figure 4.38**. Note in particular that regardless of campus rating, Asians performed similarly well. The other ethnicities showed much more variation at the different campuses. However, we can clearly see the main of effect of ethnicity in this graph: the descending order of scores is Asian, White, Hispanic, and then Black regardless of campus rating. Unless schools were targeting the different ethnicities for instructional quality, one would expect the gap between ethnicities to be insignificantly small, lending credibility to the existence of a Profiling latent trait. Lastly, the interaction between SES and campus rating showed that only non-disadvantaged students (middle and high SES) attended Exemplary schools. **Figure 4.39** shows that Other Forms of Economically Disadvantaged category (9) exhibited a departure in behavioral trend from the other categories of SES.

When taken together, there is strong circumstantial evidence that instructional quality does not matter and thus the school the students attend does

not matter. The item behavioral analysis indicates a large commonality across domains. This means that regardless of what the domain teachers may be teaching, the TAKS exam does not seem to be testing for it. Then in examining the disaggregated data, it was clear that student scores are related to the student intrinsic factors sex, ethnicity, and SES. Since student demographics influences school environment but the reverse is not true (school environment influences student demographics), it makes sense that the students matter more relative to test scores rather than school. Some might argue that parents select where to live based on school quality. This does not influence school demographics, but rather perpetuates them and is most relevant to schools that perform well and families with the means to move into the neighborhoods those schools serve. Lastly, the effect size of campus rating and its interaction effects are the smallest for all the effects indicating that the student factors matter more. Thus the claim that the TAKS exam is instructionally insensitive is sustained.

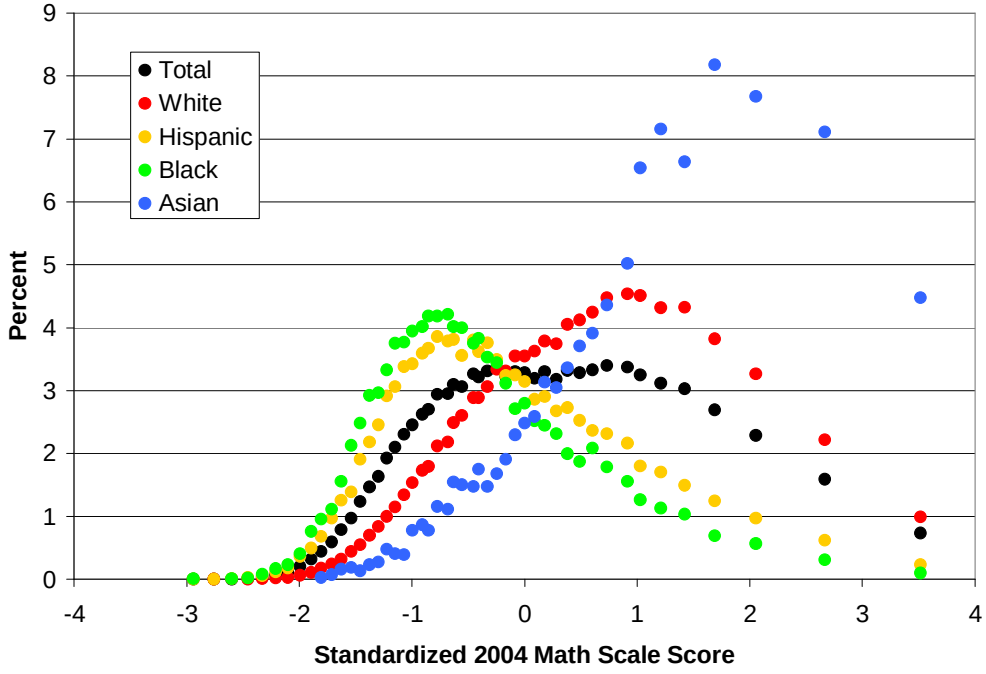

Figure 4.25 Distribution of standardized 2004 Math scale scores within each ethnicity

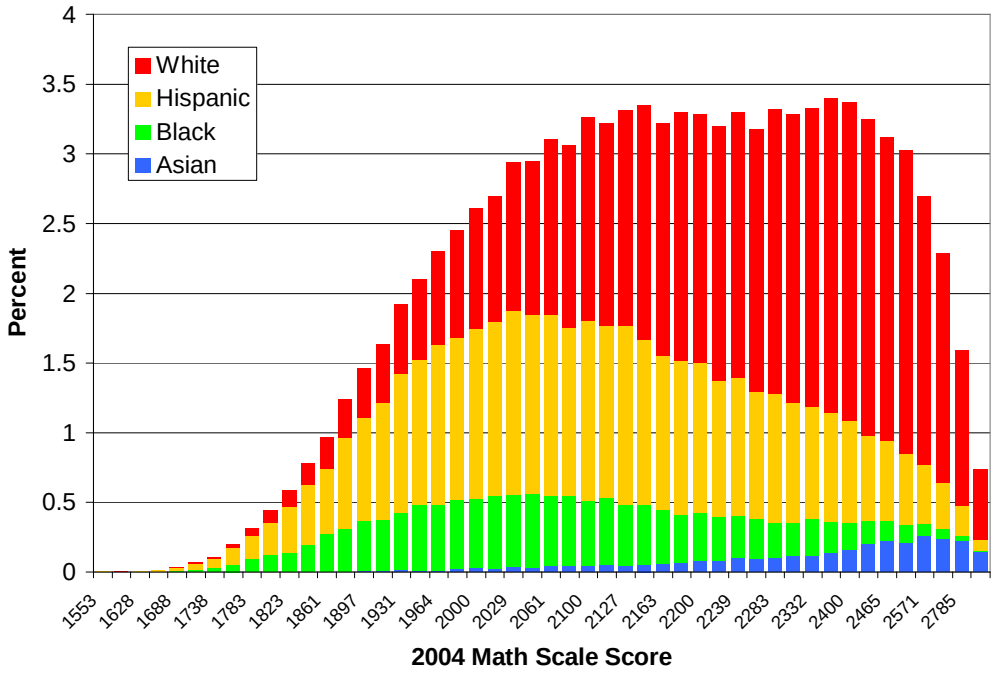

Figure 4.26 Relative contribution of each ethnicity to the total distribution of the 2004 Math scale scores

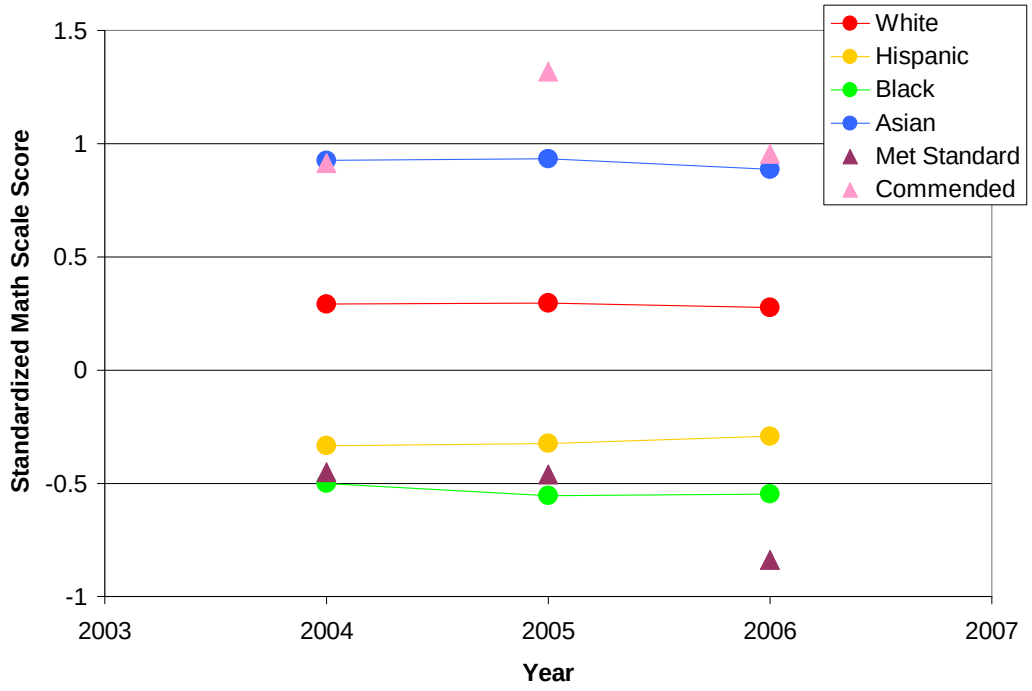

Figure 4.27 Mean standardized Math scale score by ethnicity and year

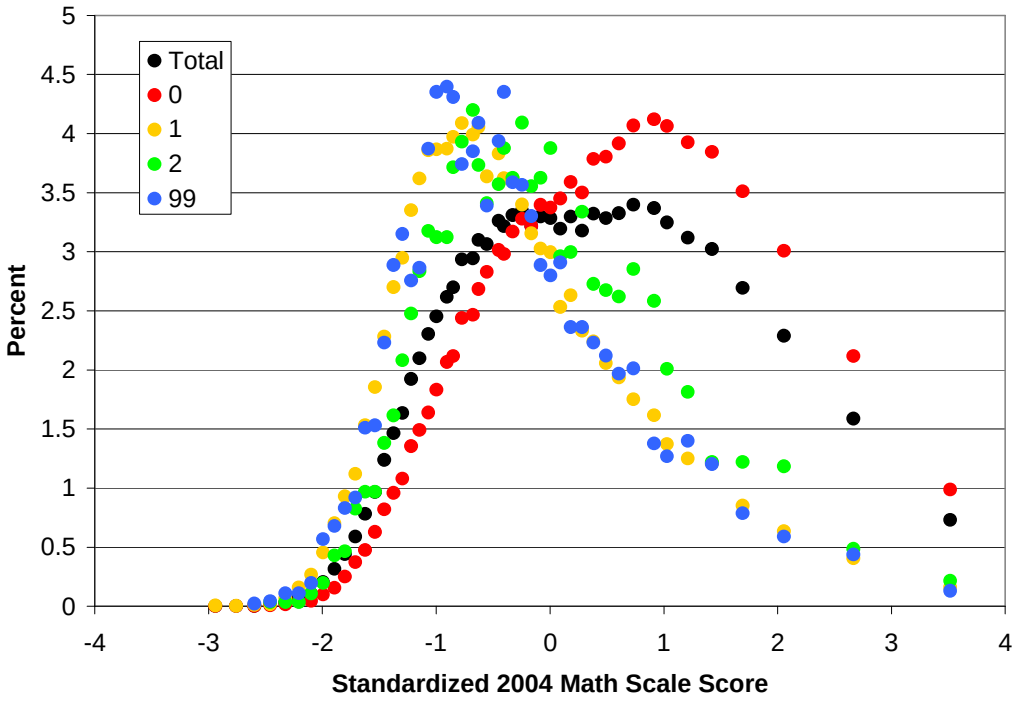

Figure 4.28 Distribution of standardized 2004 Math scale scores within each SES

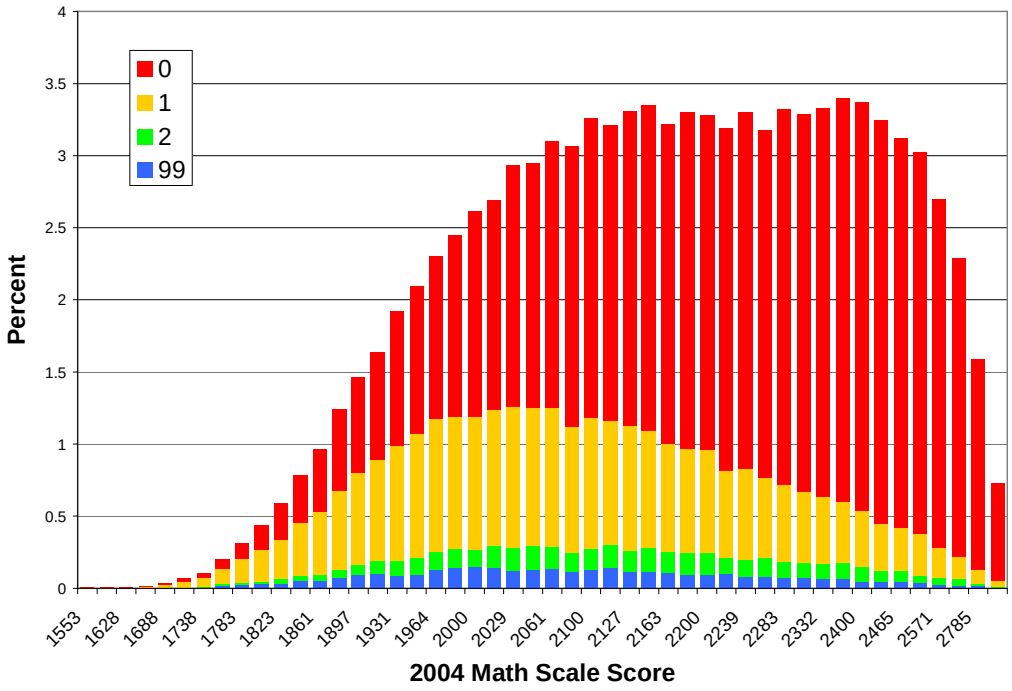

Figure 4.29 Relative contribution of each SES to the total distribution of the 2004 Math scale scores

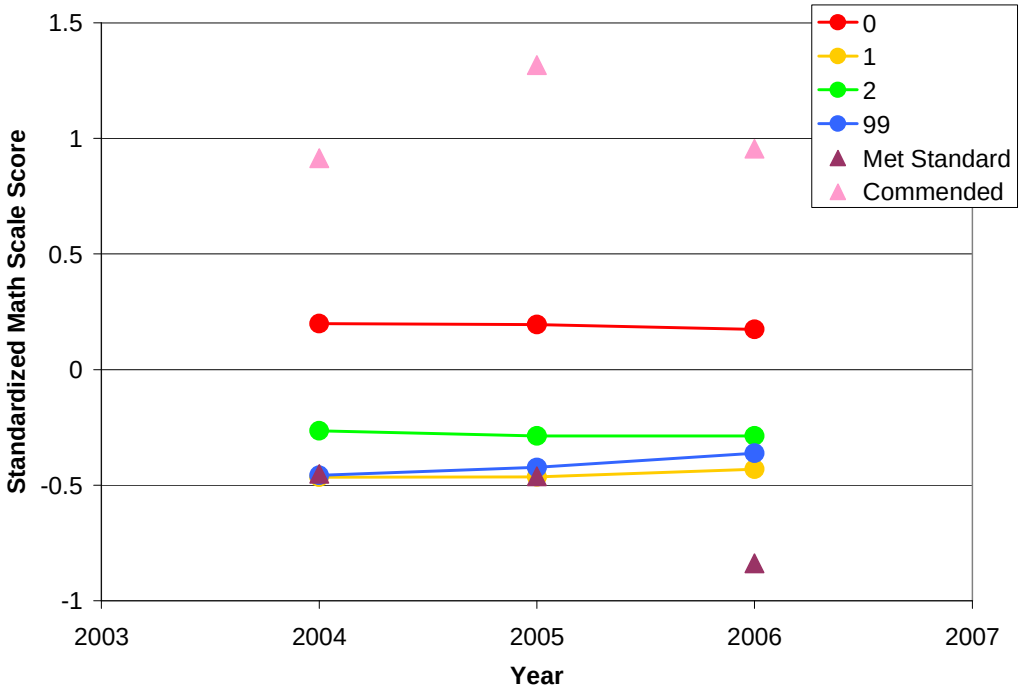

Figure 4.30 Mean standardized Math scale score by SES and year

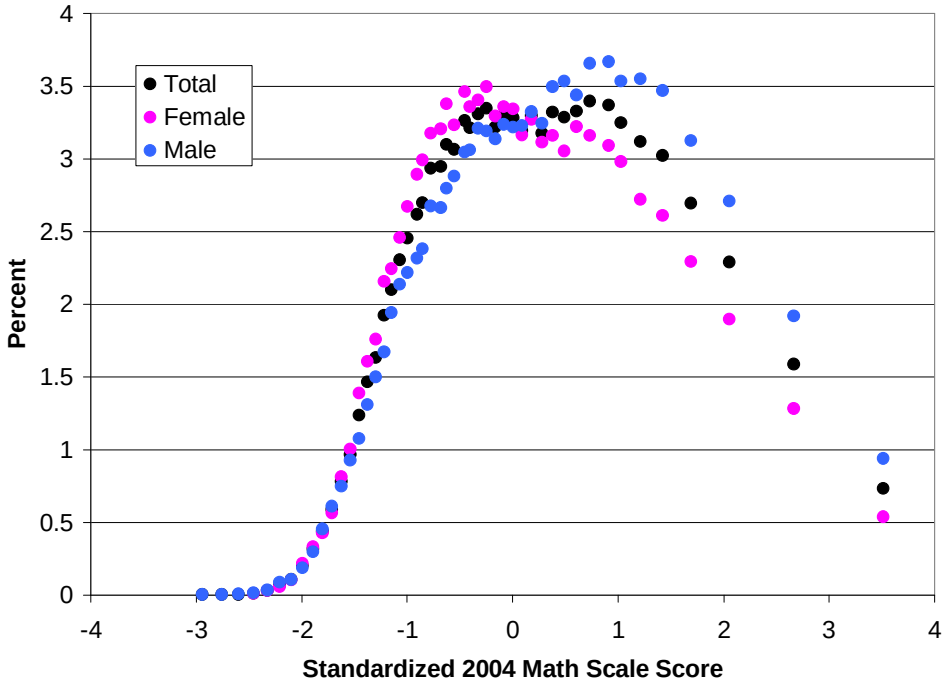

Figure 4.31 Distribution of standardized 2004 Math scale scores within each gender

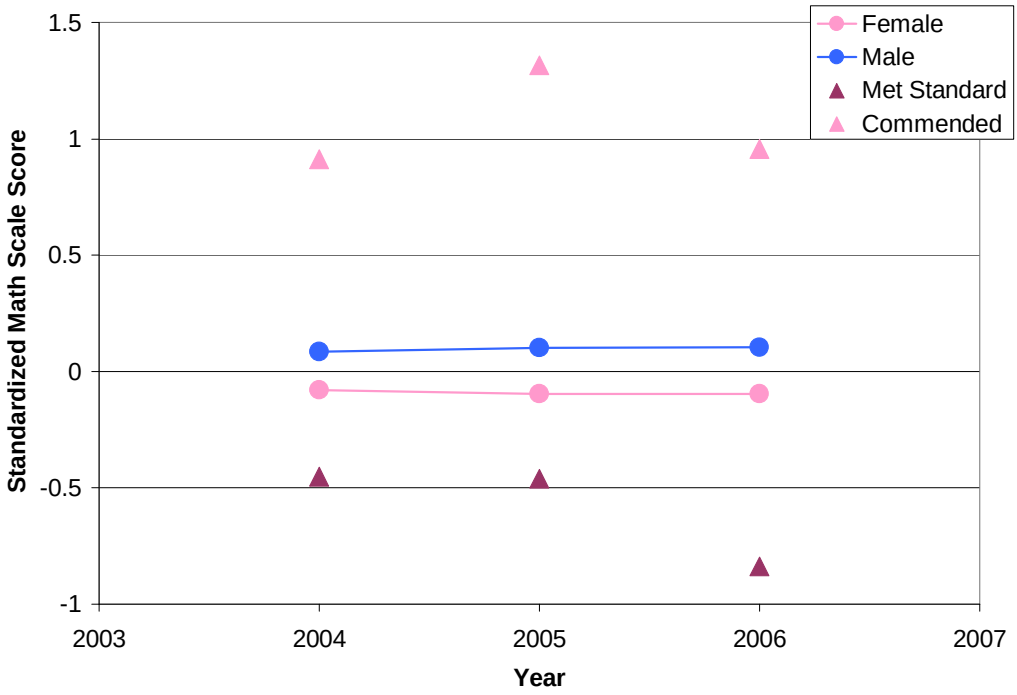

Figure 4.32 Mean standardized Math scale score by gender and year

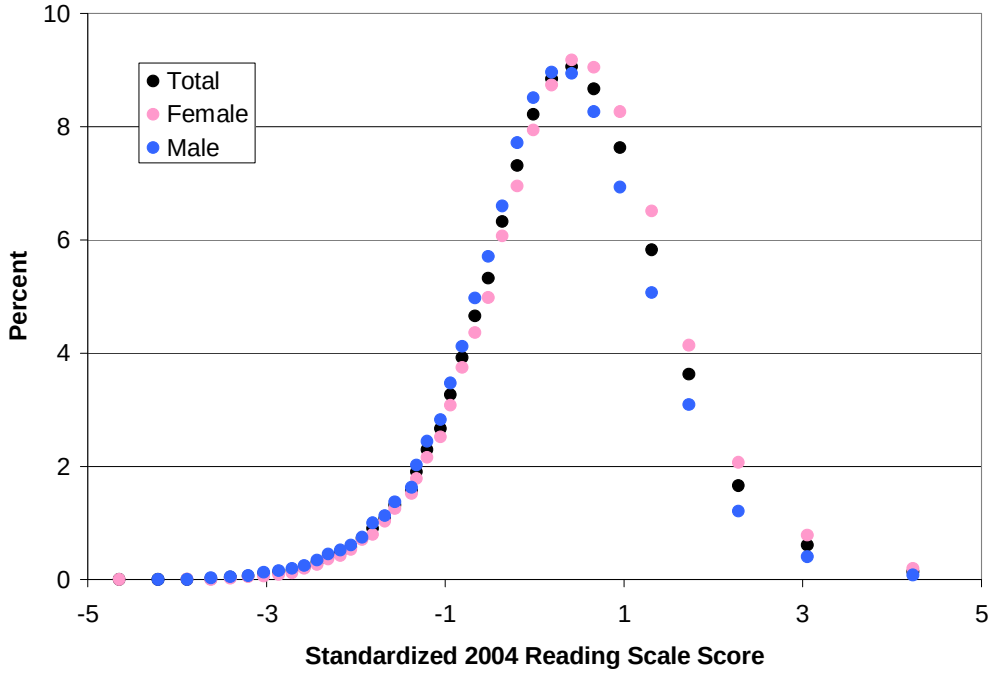

Figure 4.33 Distribution of standardized 2004 Reading scale scores within each gender

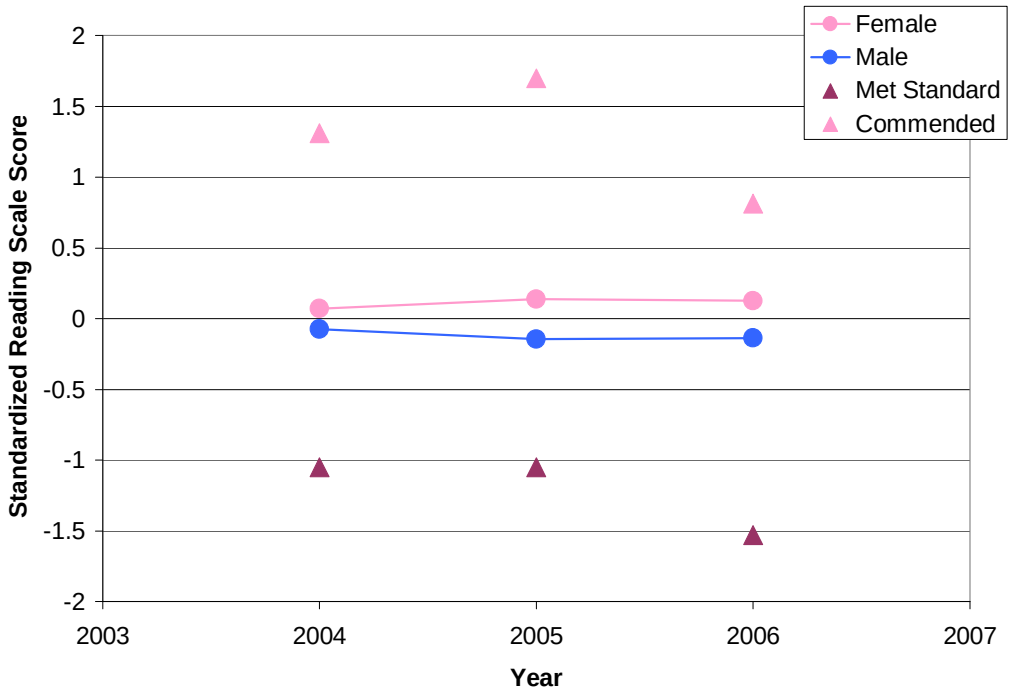

Figure 4.34 Mean standardized Reading scale score by gender and year

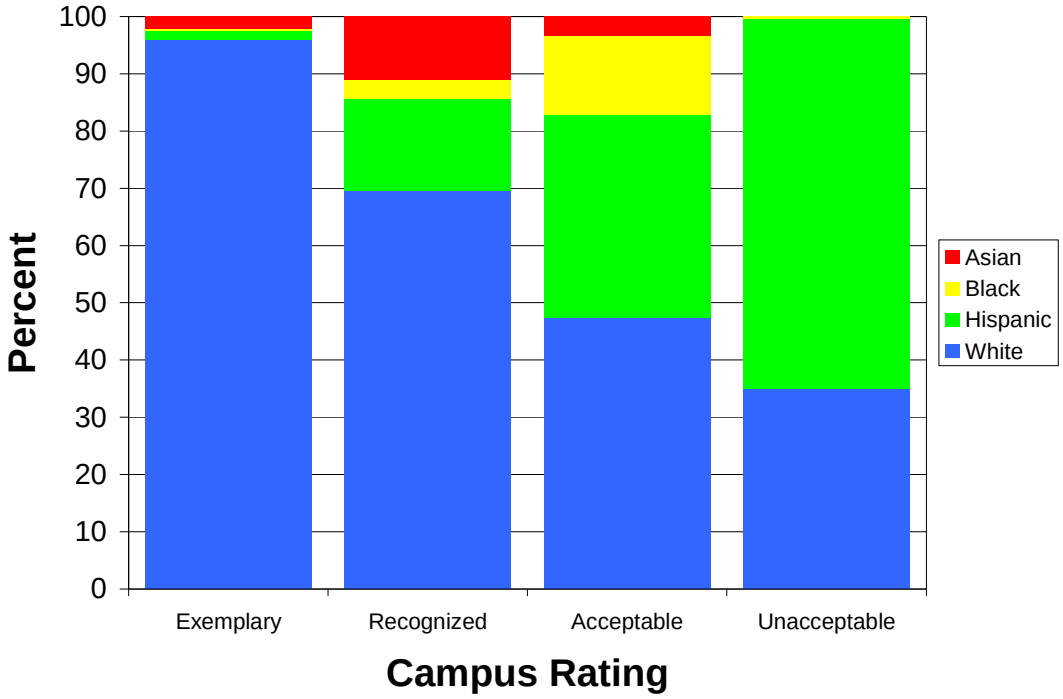

Figure 4.35 Graph of the percentage of student ethnicities at each campus rating

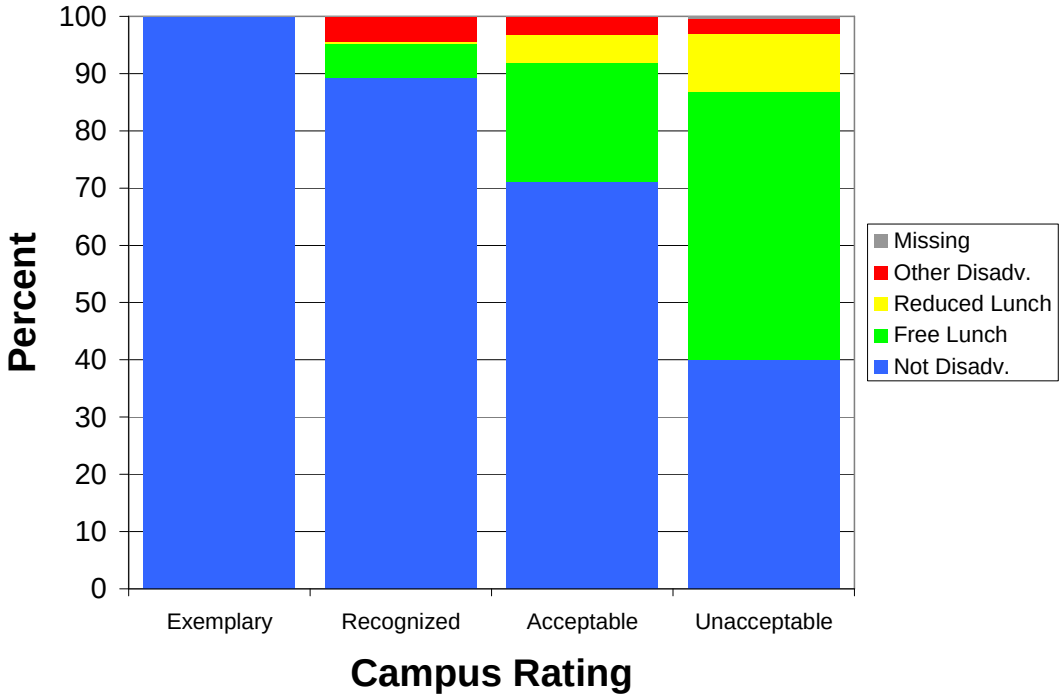

Figure 4.36 Graph of the percentage of student SES at each campus rating

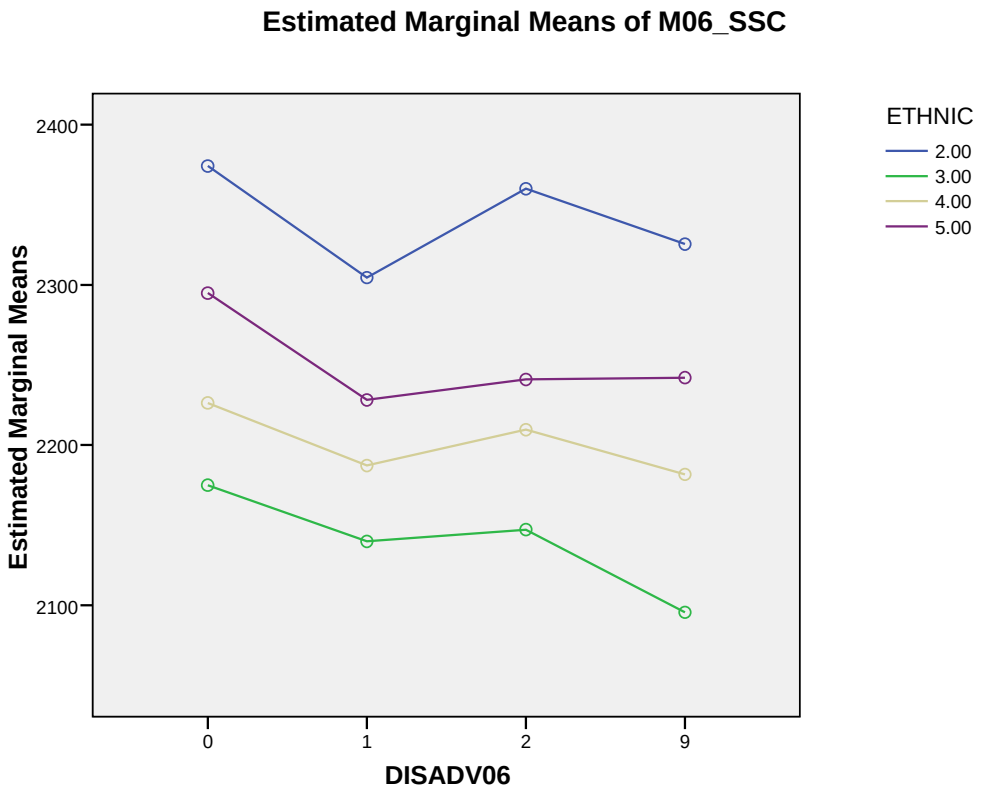

Figure 4.37 Interaction between Ethnicity and SES

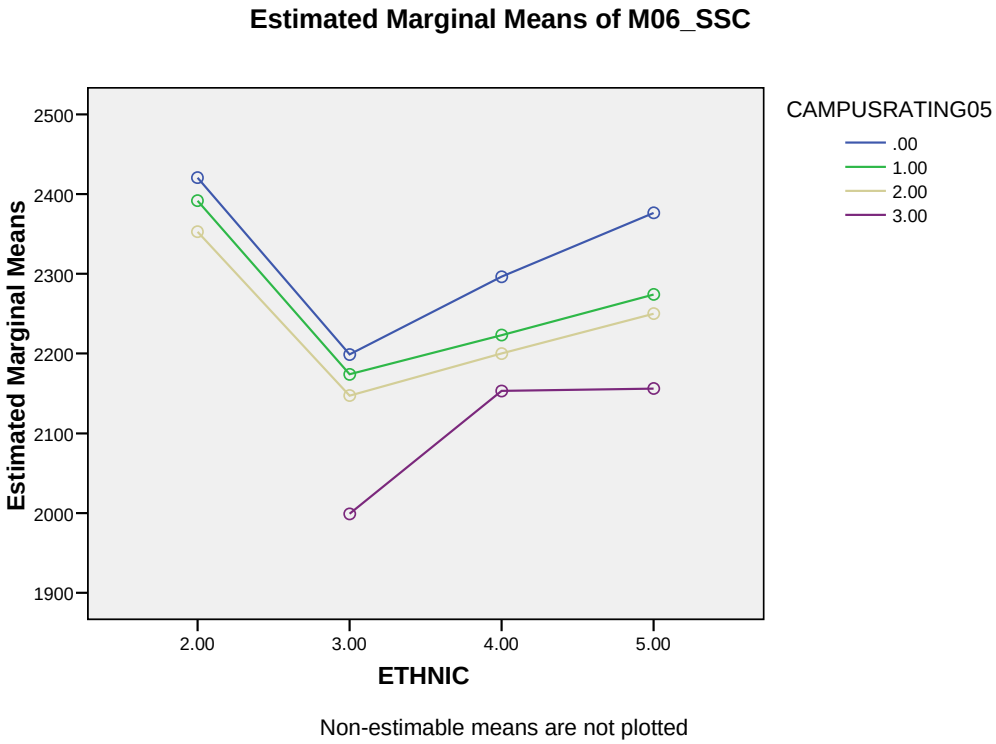

Figure 4.38 Interaction between ethnicity and campus rating

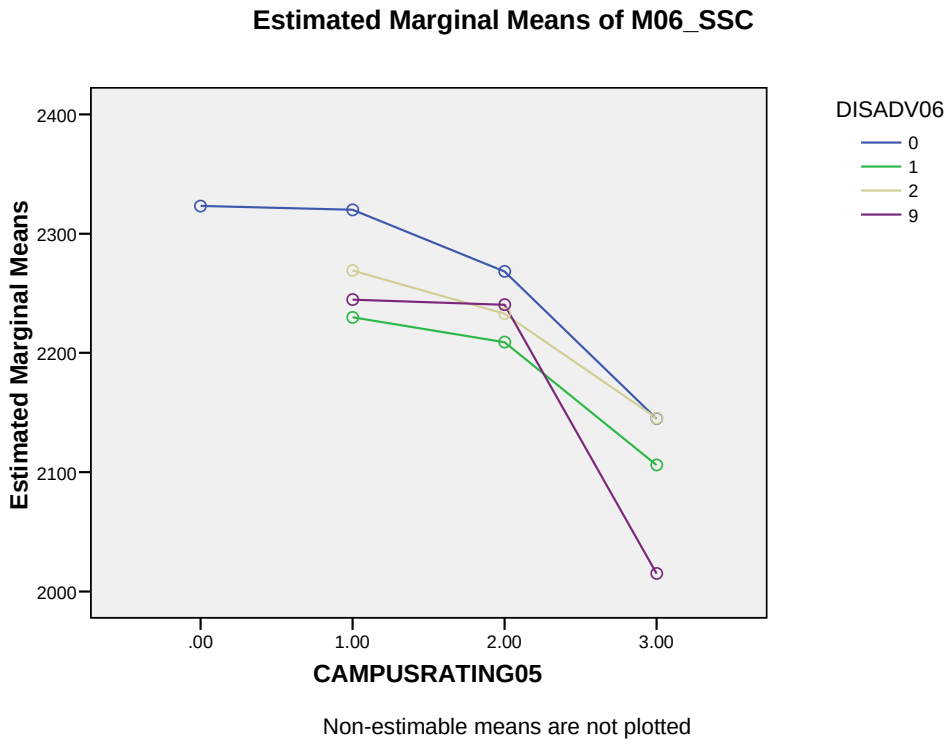

Figure 4.39 Interaction between SES and campus rating

Code	Translation
0	Not identified as economically disadvantaged
1	Eligible for free meals under the National School Lunch and Child Nutrition Program
2	Eligible for reduced-price meals under the National School Lunch and Child Nutrition Program
99	Other economic disadvantage, including: a) from a family with an annual income at or below the official federal poverty line, b) eligible for Temporary Assistance to Needy Families (TANF) or other public assistance, c) received a Pell Grant or comparable state program of need-based financial assistance, d) eligible for programs assisted under Title II of the Job Training Partnership Act (JTPA), or e) eligible for benefits under the Food Stamp Act of 1977.

Table 4.11 Codes used to identify socioeconomic status in student data (reproduced from TEA)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Effect Size	Observed Power(a)
Corrected Model	469061821(b)	33	14213995	581.979	.000		1.000
Intercept	1031829443	1	1031829443	42247	.000		1.000
SEX	20270333	1	20270333	829.950	.000	0.093	1.000
ETHNIC	1217208	3	405736	16.612	.000	0.023	1.000
DISADV06	2550617	3	850206	34.811	.000	0.033	1.000
CAMPUSRATING05	623699	3	207900	8.512	.000	0.016	.973
ETHNIC * DISADV06	4745141	9	527238	21.587	.000	0.045	1.000
ETHNIC * CAMPUSRATING05	418627	8	52328	2.143	.029	0.013	.683
DISADV06 * CAMPUSRATING05	1006402	6	167734	6.868	.000	0.021	.997
Error	2346859228	96090	24424				
Total	492198088205	96124					
Corrected Total	2815921049	96123					

a Computed using alpha = .01
b R Squared = .167 (Adjusted R Squared = .166)

Table 4.12 ANOVA results

Ethnicity	Code
White	5
Hispanic	4
Black	3
Asian	2

Campus Rating	Code
Exemplary	0
Recognized	1
Acceptable	2
Unacceptable	3

Table 4.13 Coding for ANOVA

CHAPTER 5: Computer Modeling of the TAKS Exam

In this chapter, we will explore the TAKS exam by using an agent based computer model of the underlying theoretical framework of the TAKS exam. To be able to do this would require that we first calibrate the computer model so that the scales used in the model coincide with the TAKS exam. Once the scales are calibrated, we will run a simulation to see what the data looks like if the initial conditions are similar to the real world. This will provide a baseline comparison of how a perfect execution of IRT-1Pl would look like using the TAKS exam.

Domain Linkage Set Up

Based on the SEM analysis (**Figure 4.24**) of the data from the real world, it is known that the domain latent traits are related to LTP. It is unknown, however, what value of linkage should be used to produce the shared variance between the domain latent traits and LTP using **Equation 3.1**. Therefore, a series of simulations were run on the model to determine what linkage values to LTP were needed for each domain. **Figure 5.1** shows the graphical results as well as the fit lines and their respective equations. With this data, it was possible to calculate the linkage values needed to simulate the correlation values between LTP and the domain latent traits as found in the SEM of the real world data analysis. These values are shown in **Table 5.1**. It should be noted that while fitting a quintic equation to **Figure 5.1** is statistically inappropriate, it afforded an easy way to interpolate values for the purpose of inputting into the computer model to derive **Table 5.1**.

Linkage values in no way affect the year to year intra-domain correlations as well as the correlation of each section to the domain latent traits of the turtles. These correlations are a result of the set of items on the exam and their interaction with the turtles' θ values. The linkage values only affect the amount of correlation between the domain latent traits and LTP as well as inter-domain correlations. This will be clearly demonstrated when comparing the 0 and 100 link runs later. A different interpretation of the linkage values is that the values represent the average amount of "domain content" and "profiling content" in each

item of that domain. With this interpretation, it can be seen that items test mostly for profiling and not very much domain content based on **Table 5.1**. This interpretation is only possible due to the *a priori* establishment of the domain latent traits in the turtles, demonstrating the utility of computer modeling.

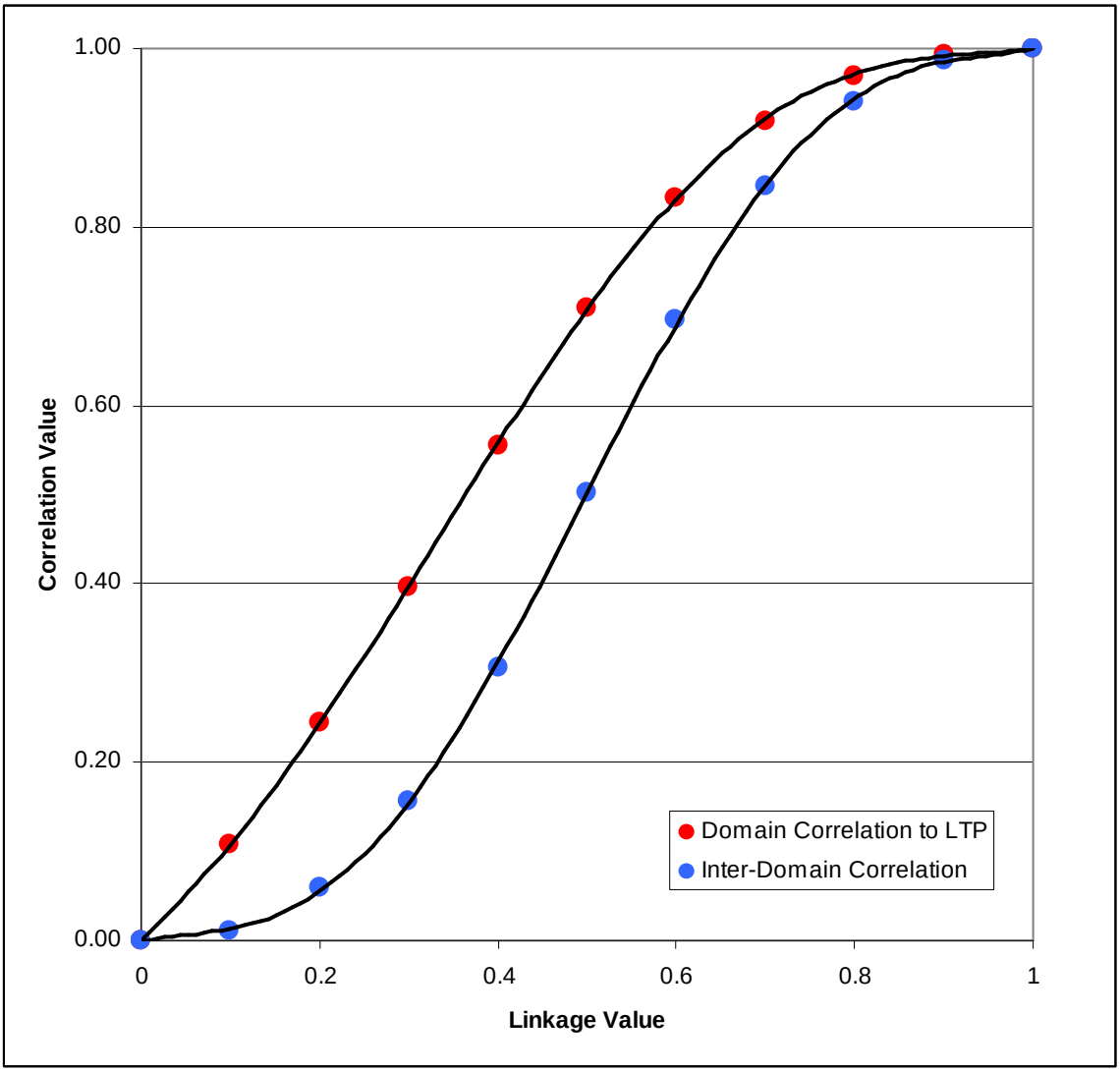

Figure 5.1 Dependence of Correlation Value to Linkage value
(Domain Correlation to LTP Fit Line: $y = 2.616x^5 – 4.7935x^4 + 0.8197x^3 + 1.4266x^2 + 0.9316x + 0.00009$, $R^2 = 1.00$;
Inter-Domain Correlation Fit Line: $y = 9.1158x^5 - 22.692x^4 + 16.484x^3 – 2.1131x^2 + 0.207x – 0.0007$, $R^2 = 0.999$)

Domain	Linkage Value	Domain Correlation to LTP
Reading	0.622	0.855
Math	0.665	0.894
History	0.744	0.948
Science	1.000	1.000

Table 5.1 Linkage values needed to have SEM domain to Profiling latent trait correlation values

Real World Population Initial Conditions

Just as it was necessary to determine the Linkage values needed to simulate the correlation values between the domains latent traits and LTP, it is necessary to determine what the initial distribution of the domain latent trait values are if a simulation of the real world is to be done. This is because the scales used on the TAKS exam were generated independently and thus have different midpoints and units of measurement and so the model must be calibrated to reflect these differences. To do this, the Analytic dataset raw scores in each domain were summed for all students and the descriptive statistics determined. Note that since the Analytic dataset Reading raw scores contain short answer and/or essay items, the Reading raw scores were re-determined based only on the multiple choice response set. **Table 5.2** shows these descriptive statistics for each domain. A series of simulations were run with varying Population Means but under constant Population Standard Deviation = 1. The results of these runs are shown on **Figure 5.2.** Then using the Fit Lines for each domain and the mean of each domain from the Analytic dataset, the Population Means were determined and shown in **Table 5.3** along with the resultant domain raw score means and standard deviations. Afterwards, a series of simulations were run using these new means but under varying Population Standard Deviations. The results of these runs are shown on **Figure 5.3**. Then using the Fit Lines for each domain and the domain standard deviations from the Analytic dataset, the Population Standard Deviation was determined and shown

in **Table 5.4** along with the resultant Domain raw score means and standard deviations. This process was repeated until the model yielded domain means and standard deviations close to those found in the Analytic dataset which took seven iterations. The determined values are shown in **Table 5.5**. Once again, fitting quartic equations to **Figures 5.2** and **5.3** is statistically inappropriate, but the goal is to interpolate values accurately, not perform a statistical analysis to draw conclusions. These mean and standard deviation values establish the baseline latent trait distribution of the Analytic dataset for the model. It is interesting to note that the relative ability level of the turtles in each domain matches the ease with which the Analytic dataset students perceived those domains on the TAKS exam. Furthermore, the positive values again indicate that overall, the mean student ability level was higher than the mean difficulty of the TAKS.

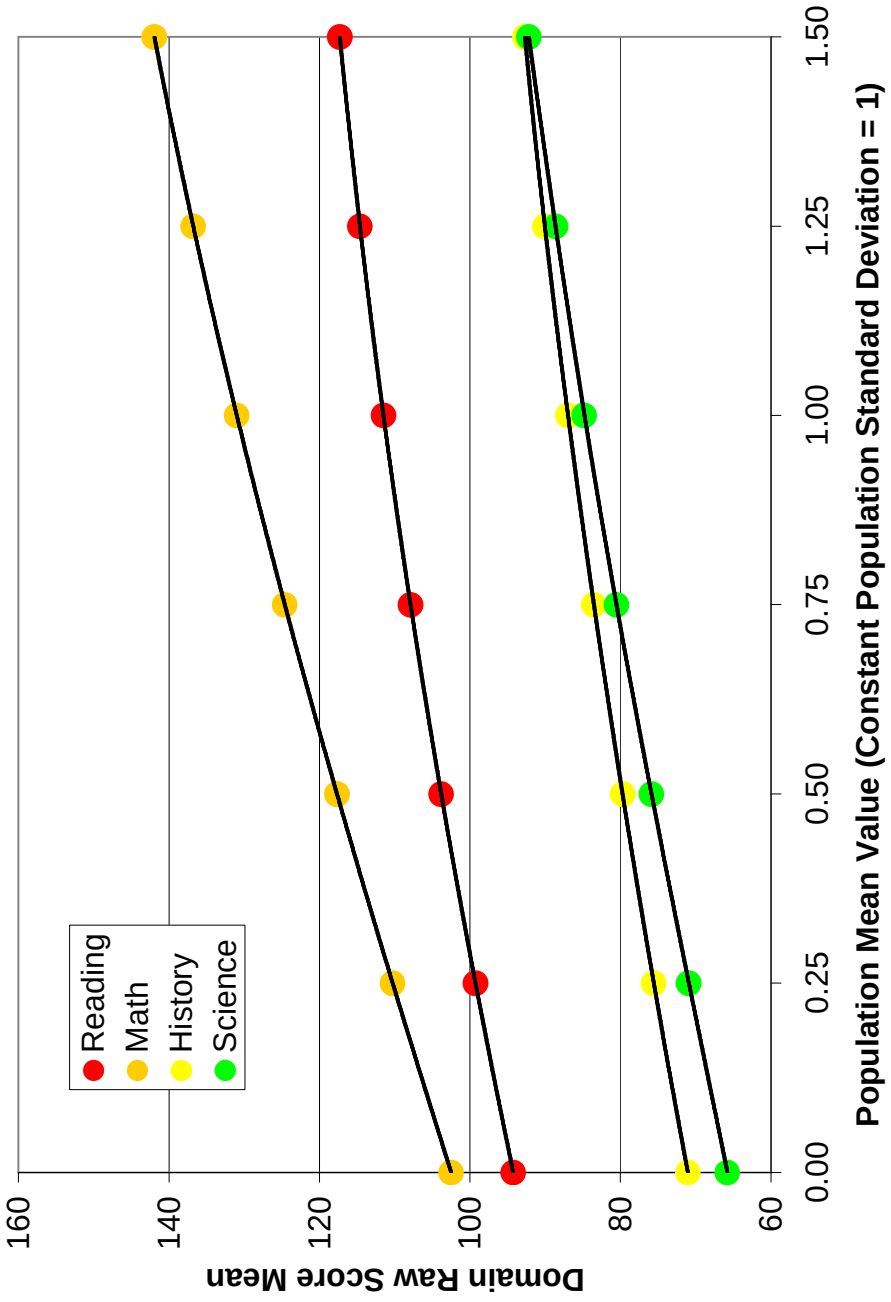

Figure 5.2 Domain Raw Score Mean as a Function of the Population Mean under Constant Standard Deviation = 1
(Reading Fit Line: $y = 0.232x^4 - 0.6686x^3 - 4.1395x^2 + 20.855x + 94.212$, $R^2 = 1$
Math Fit Line: $y = 0.2027x^4 - 1.2593x^3 - 2.2131x^2 + 31.764x + 102.49$, $R^2 = 1$
History Fit Line: $y = -0.3378x^4 + 0.8918x^3 - 4.6316x^2 + 19.02x + 71.032$, $R^2 = 1$
Science Fit Line: $y = 1.0098x^4 - 4.4473x^3 + 1.0346x^2 + 20.39x + 65.792$, $R^2 = 1$)

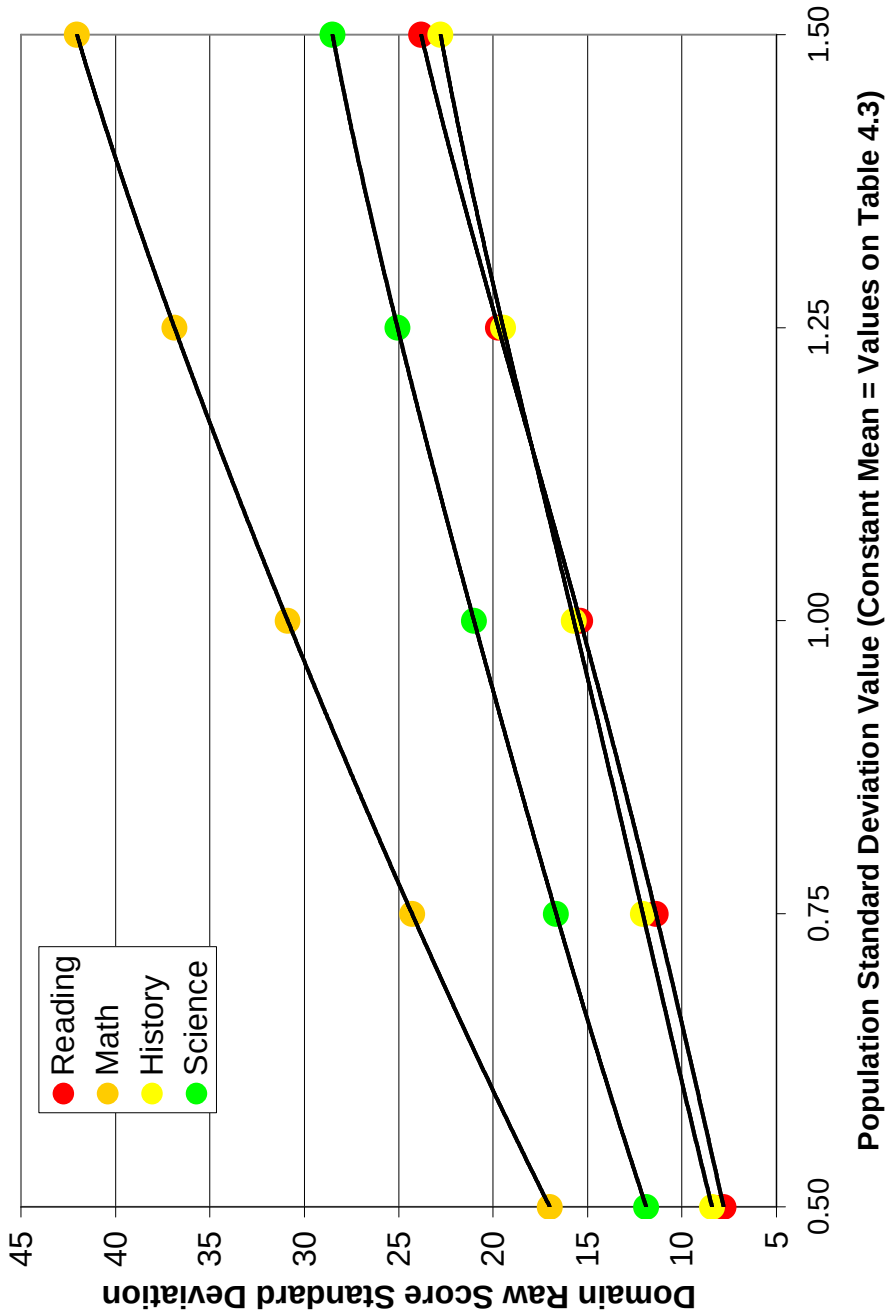

Figure 5.3 Domain Raw Score SD as a Function of the Population SD under Constant Mean from Fit Lines in *Table 5.3*
(Reading Fit Line: $y = -5.4683x^4 + 18.268x^3 - 18.967x^2 + 21.921x - 0.3686$, $R^2 = 1$
Math Fit Line: $y = y = -4.8122x^4 + 14.61x^3 - 25.44x^2 + 47.508x - 1.9849$, $R^2 = 1$
History Fit Line: $y = -6.3319x^4 + 24.192x^3 - 30.653x^2 + 31.839x - 2.3606$, $R^2 = 1$
Science Fit Line: $-4.3255x^4 + 16.468x^3 - 25.378x^2 + 35.679x - 1.3911$, $R^2 = 1$)

Domain	Minimum	Maximum	Mean	Std. Deviation
Reading Raw Score	27	129	109.95	12.60
Math Raw Score	32	168	115.43	28.49
History Raw Score	19	105	83.08	14.78
Science Raw Score	18	110	74.42	17.46

Table 5.2 Real world Domain Raw Score Descriptive Statistics (N = 69,570)

Domain	Population Mean Values	Resultant Domain Raw Score Mean	Resultant Domain Raw Score SD
Reading	0.889	110.0052	15.3854
Math	0.422	115.3706	30.8804
History	0.720	84.942	15.7251
Science	0.375	74.4157	21.0125

Table 5.3 Resulting Domain Statistics after First Iteration of Model Using Constant Standard Deviation = 1

Domain	Population SD Values	Resultant Domain Raw Score Mean	Resultant Domain Raw Score SD
Reading	0.830	111.2256	12.6218
Math	0.907	116.2378	28.4066
History	0.936	84.5282	14.8470
Science	0.813	74.3032	17.8565

Table 5.4 Resulting Domain Statistics after First Iteration of Model Using Population Mean Values from *Table 5.3*

Domain	Population Mean	Population SD	Resultant Raw Score Mean	Resultant Raw Score SD
Reading	0.777	0.780	109.91	12.65
Math	0.395	0.898	115.46	28.46
History	0.681	0.915	83.09	14.77
Science	0.325	0.780	74.43	17.48

Table 5.5 Population Means and Standard Deviations of the Latent Traits Needed to Simulate Real world data

Sample Runs

Zero Link Run

In order to see what the results of the TAKS exam would be if each domain were perfectly orthogonal to each other, a zero link simulation was run. Note that in this run, the population values for domain latent trait means and standard deviations were set to Real world values as determined previously. Only the domain linkage values are set to zero. **Table 5.6** shows the correlation matrix for the turtle θ values after turtles who scored perfectly in any section were removed just like in the IRT Comparison dataset.

There are a few things to note on this table. Only intra-domain θ values have any level of significant correlation. Since the model creates orthogonal domain latent trait values, this is to be expected. However, the intra-domain correlation values are very close to that of the intra-domain correlation values from the real world (see **Table 5.8**). The intra-domain correlations are purely a result of the turtles' θ values and the set of item b-values in each section. Also, the scores on each section within a domain are better correlated to their respective domain latent traits than they are with each other. These correlations are less than perfect due to the error of estimation and represent the maximal ability of each section to estimate turtle θ values for each domain latent trait. Recall that the ability of an IRT exam to accurately and precisely estimate θ values is directly related to the number of items and the interval of difficulties between items on that exam. That is why Reading with the fewest number of

items and being the easiest domain, has the lowest section to domain latent trait correlation values of all the domains.

Item analysis of the items in **Table 4.6** clearly shows what the IRFs would look like if the domains were truly orthogonal. **Figure 5.4** through **5.7** show these IRFs. Once again, the stochasticity seen at the low end is due to small numbers of turtles with such low θ values. Only when the item domain and scale match does a clear IRF present itself. When the item domain and scale do not match, a horizontal line at around y = 0.80 forms since this is the y-intercept for the item with b-values around -0.90.

In an ideal world, we would want to select for items that are orthogonal to each other in terms of the domain being test. This ensures that we are testing students on domain specific knowledge and skills only. However, this can never truly happen. To be able to read, decode, and respond to an item requires so many different skills, that there must be some level of shared commonalities across domains. In the next run, we see what would happen if the shared commonalities are perfect with no actual domain specific content being measured.

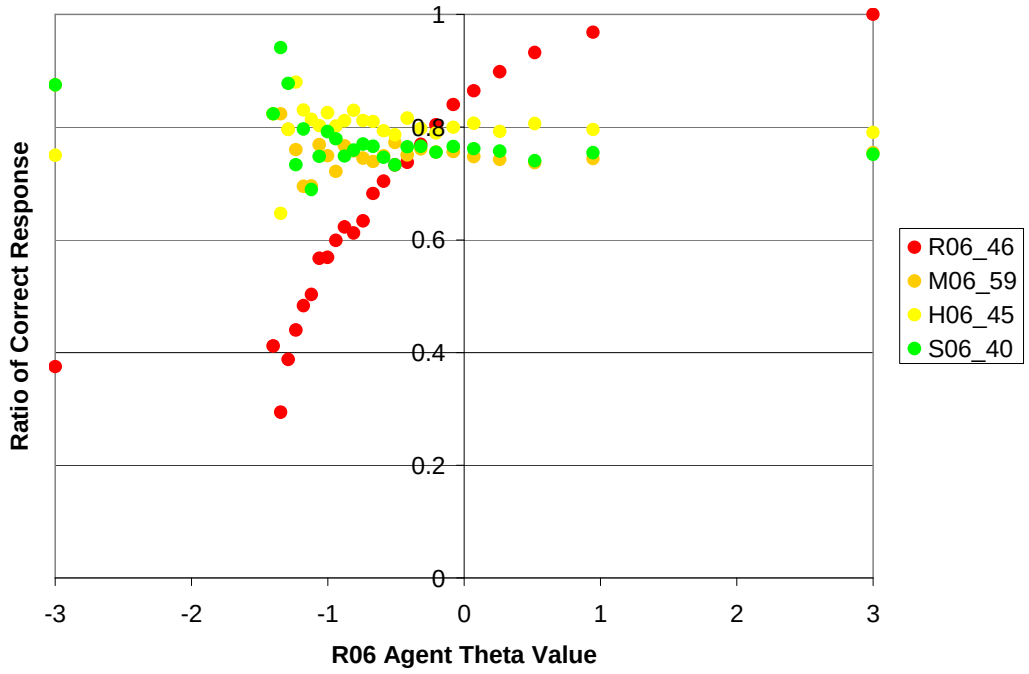

Figure 5.4 IRFs of the different domain items using R06 Agent θ for a zero domain linkage run

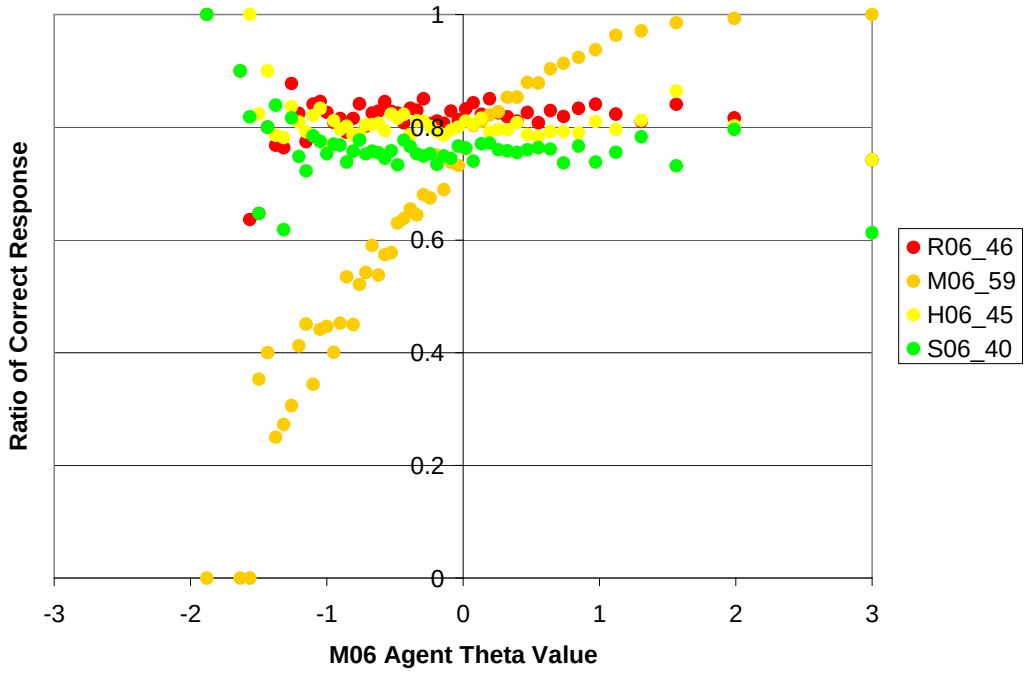

Figure 5.5 IRFs of the different domain items using M06 Agent θ for a zero domain linkage run

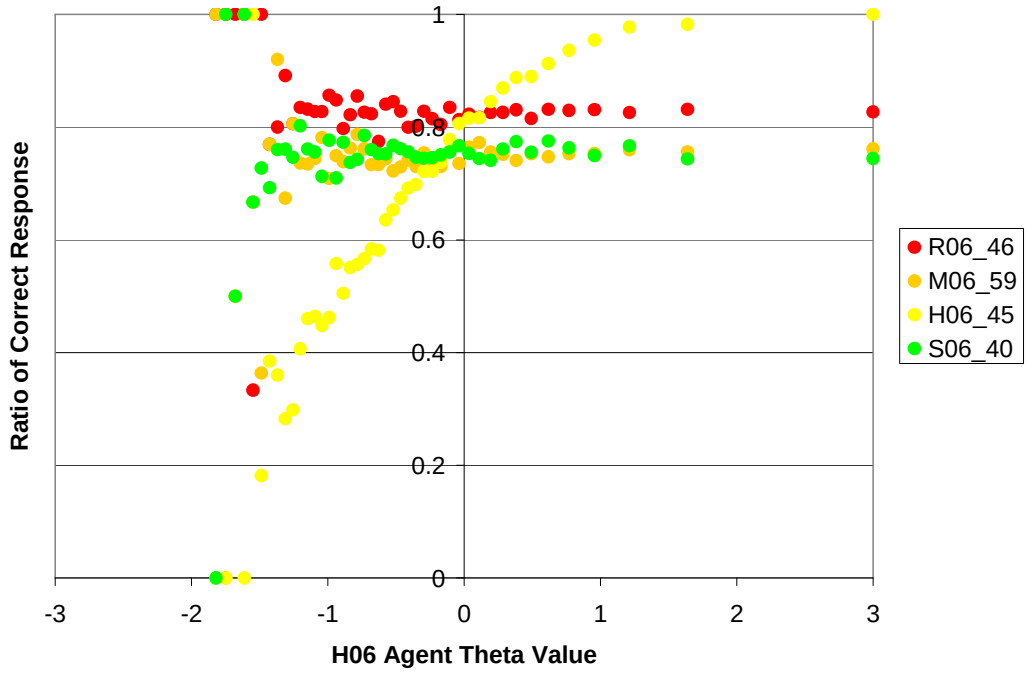

Figure 5.6 IRFs of the different domain items using H06 Agent θ for a zero domain linkage run

Figure 5.7 IRFs of the different domain items using S06 Agent θ for a zero domain linkage run

	R04 Agent θ	R05 Agent θ	R06 Agent θ	LTR	M04 Agent θ	M05 Agent θ	M06 Agent θ	LTM	H05 Agent θ	H06 Agent θ	LTH	S05 Agent θ	S06 Agent θ	LTS	LTP
R04 Agent θ	1.00														
R05 Agent θ	**0.71**	1.00													
R06 Agent θ	**0.69**	**0.70**	1.00												
LTR	0.83	0.85	0.83	1.00											
M04 Agent θ	0.00	0.00	0.01	0.00	1.00										
M05 Agent θ	0.00	0.00	0.01	0.00	**0.87**	1.00									
M06 Agent θ	0.00	0.00	0.01	0.00	**0.87**	**0.88**	1.00								
LTM	0.00	0.00	0.01	0.01	0.93	0.94	0.94	1.00							
H05 Agent θ	-0.01	-0.01	-0.02	-0.01	0.00	0.00	0.00	0.00	1.00						
H06 Agent θ	-0.01	-0.01	-0.02	-0.02	0.00	0.00	0.00	0.00	**0.83**	1.00					
LTH	-0.01	-0.01	-0.02	-0.01	0.00	0.00	0.00	0.00	0.91	0.91	1.00				
S05 Agent θ	0.01	0.01	0.01	0.01	-0.01	-0.01	-0.01	-0.01	0.01	0.01	0.01	1.00			
S06 Agent θ	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.85**	1.00		
LTS	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.01	0.00	0.00	0.92	0.92	1.00	
LTP	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.00	0.00	1.00

Table 5.6 Correlation Matrix of Agent θ values for a Zero Domain Linkage Run (N = 26,237)

100 Link Run

The 100 Link run was also performed with all parameters set the to the same values as the zero link run except for the domain linkage values being set to 100. All domains are now perfectly correlated with each other and determined solely by LTP. **Table 5.7** is the correlation matrix of the θ values after agents who scored perfectly in any section were removed. The intra-domain correlations are close to that seen in **Table 5.6** since they are not affected by domain linkage. Minor differences are due to the randomness that was built into the computer model. In this run, the inter-domain correlations are also significant. The values of the inter-domain correlations represent the highest amount of correlation possible between the different sections based on the set of b-values in each section. In other words, even though the domains are perfectly linked, the inter- and intra-domain correlations are not 100 because each section cannot perfectly measure the turtles even if the underlying latent traits are perfectly correlated. As the theoretical upper limit in terms of correlations, the inter-domain correlations are also higher than those of the IRT Comparison dataset in **Table 5.8**. Still they are close in value which would indicate that the domains share a large amount of variance as was seen in the SEM of the real world data. **Figures 5.8** through **5.11** show the IRFs for the different items in this run. Not surprisingly, all of them have clean IRFs that are about the same. The differences are due to the fact that each item domain may have a b-value ~-0.90, but none of them are actually the exact same value. Furthermore, the rescaling of each domain latent trait value to

simulate the scales of each domain in the real world caused the items to behave slightly different when using a different scale than the item.

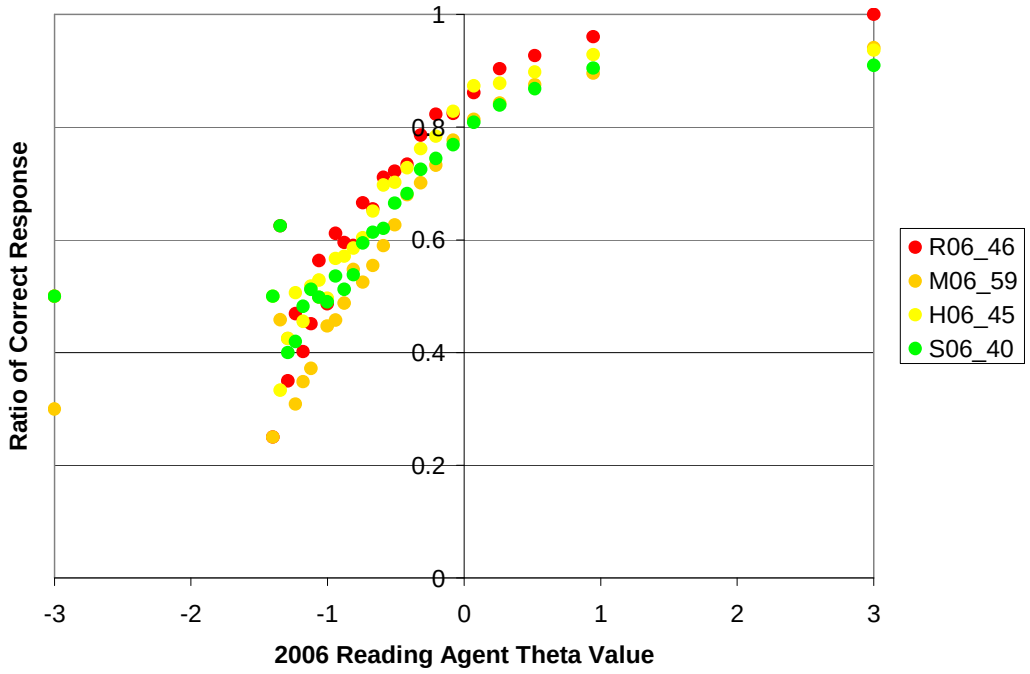

Figure 5.8 IRFs of the different domain items using R06 Agent θ for a 100 domain linkage run

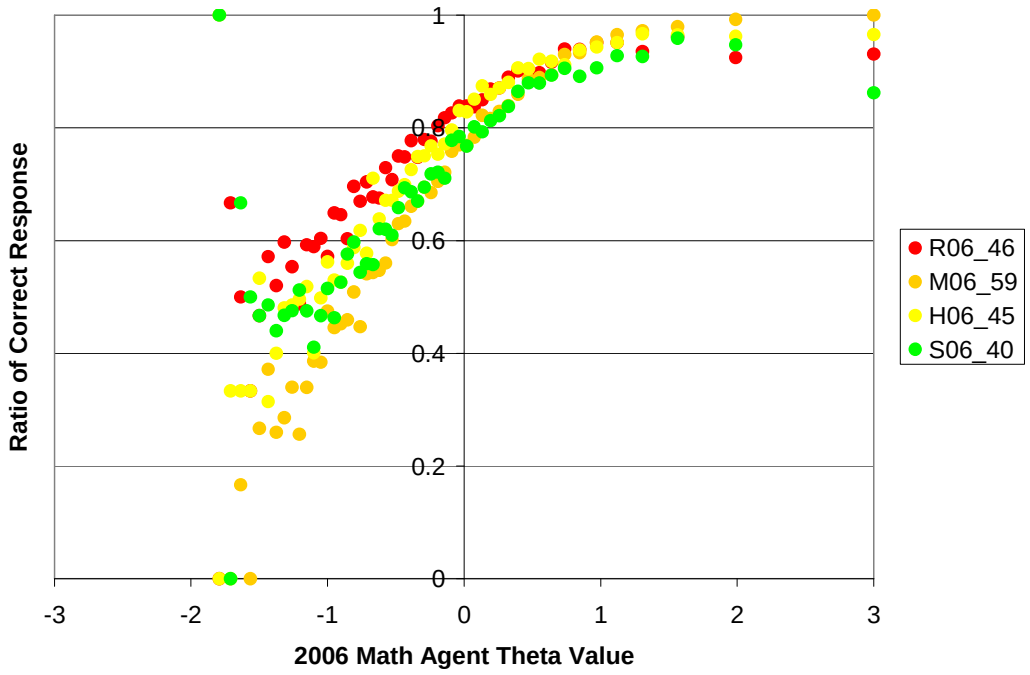

Figure 5.9 IRFs of the different domain items using M06 Agent θ for a 100 domain linkage run

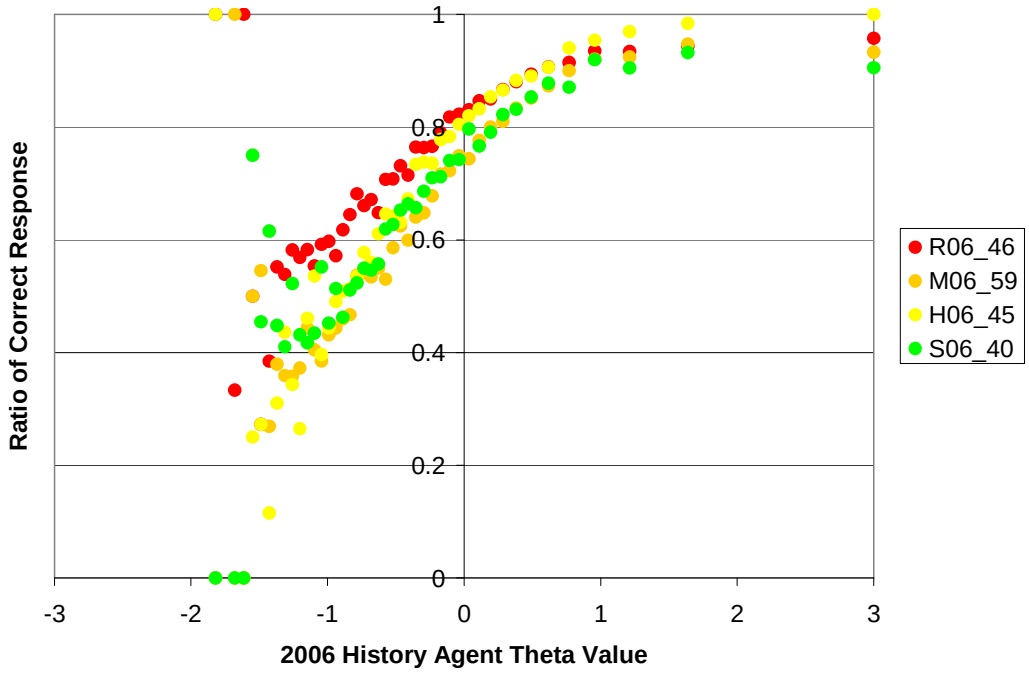

Figure 5.10 IRFs of the different domain items using H06 Agent θ for a 100 domain linkage run

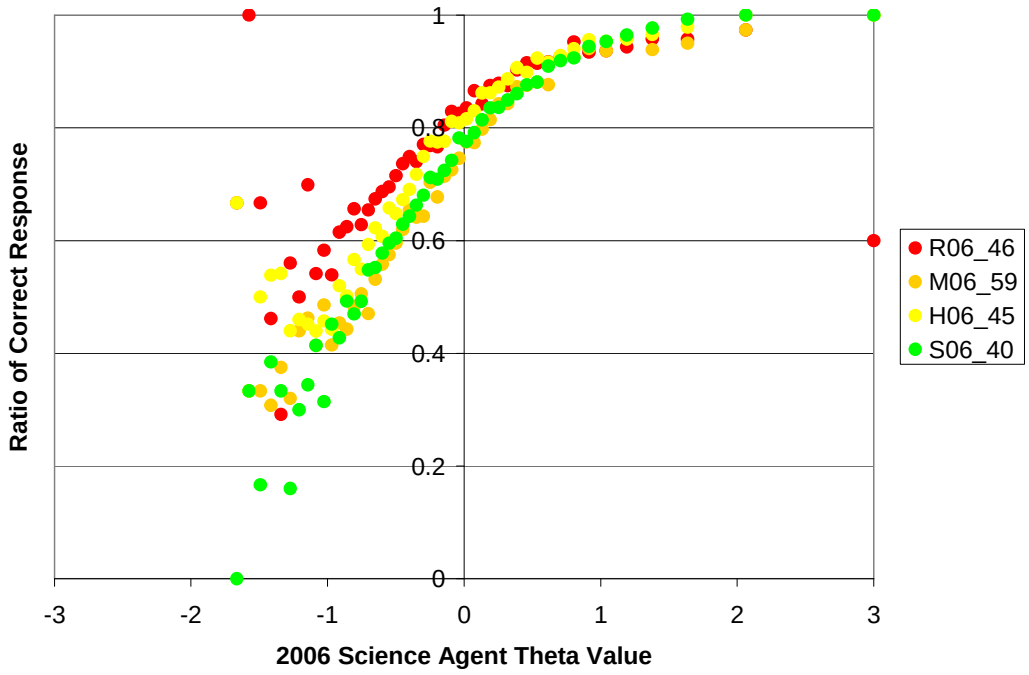

Figure 5.11 IRFs of the different domain items using R06 Agent θ for a 100 domain linkage run

	R04 Agent θ	R05 Agent θ	R06 Agent θ	LTR	M04 Agent θ	M05 Agent θ	M06 Agent θ	LTM	H05 Agent θ	H06 Agent θ	LTH	S05 Agent θ	S06 Agent θ	LTS	LTP
R04 Agent θ	1.00														
R05 Agent θ	**0.70**	1.00													
R06 Agent θ	**0.68**	**0.70**	1.00												
LTR	0.83	0.85	0.82	1.00											
M04 Agent θ	0.77	0.78	0.76	0.93	1.00										
M05 Agent θ	0.77	0.79	0.77	0.93	**0.86**	1.00									
M06 Agent θ	0.77	0.79	0.77	0.93	**0.87**	**0.87**	1.00								
LTM	0.83	0.85	0.82	1.00	0.93	0.93	0.93	1.00							
H05 Agent θ	0.75	0.76	0.75	0.90	0.84	0.84	0.84	0.90	1.00						
H06 Agent θ	0.75	0.77	0.75	0.91	0.84	0.85	0.85	0.91	**0.82**	1.00					
LTH	0.83	0.85	0.82	1.00	0.93	0.93	0.93	1.00	0.90	0.91	1.00				
S05 Agent θ	0.76	0.77	0.76	0.92	0.85	0.85	0.86	0.92	0.83	0.83	0.92	1.00			
S06 Agent θ	0.76	0.77	0.75	0.92	0.85	0.86	0.85	0.92	0.83	0.83	0.92	**0.84**	1.00		
LTS	0.83	0.85	0.82	1.00	0.93	0.93	0.93	1.00	0.90	0.91	1.00	0.92	0.92	1.00	
LTP	0.83	0.85	0.82	1.00	0.93	0.93	0.93	1.00	0.90	0.91	1.00	0.92	0.92	1.00	1.00

Table 5.7 Correlation Matrix of Agent θ values for a 100 Domain Linkage Run (N = 26,725)

Simulation Run

So far the model has only been run under the extreme condition of no or full domain linkage. In order to see how the model behaves when compared to real world settings, the same population values for the domain latent traits were used as before as well the domain linkage values from the real world SEM (**Table 5.1**). Recall that using these values merely set up the initial conditions of the model population. How the model population behaves is based only on the mathematical equations as set forth by IRT-1PL as was shown in the 0 and 100 link run. This simulation run will provide a baseline comparison of how well the IRT-1PL framework was implemented between real world and theoretical model. The same analyses were used for the model population that was used for the real world students.

Student Behavioral Trends

For the following student behavioral trends analyses, all turtles who made a perfect score on any section were removed since we need only turtles that were measured by IRT. This left an N = 26,657 turtles from the original 30,000 in the model. **Table 5.8** is the correlation matrix of the simulated run. While all correlation values tend to be higher than the values found in **Table 4.8**, they are also very close. **Figures 5.12** through **5.15** show the graphs of how turtles performed between 2004-2005 and 2005-2006 Math TAKS domain and are comparably to **Figures 4.1** through **4.4**. Note the similar distribution as well as

similar fit lines and R^2 values. Turtles tended to perform similarly across years and maintaining their relative rank ordering, just as real students do. Since the turtles obviously do not change between testing, any differences in scores from year to year are purely a result of random error.

Figures 5.16 and **5.17** shows the graph of the difference in turtle scores between 2005 and 2004 and the difference in scores between 2006 and 2005, and is comparable to **Figures 4.5** and **4.6**. Continuing in the same vein as before, the fit line and R^2 are close to real world values. It is clear from **Figure 5.16** that the scatter due to random error in the computer model is actually less than that observed from the real word data. One would not expect it to be otherwise since there is no such thing as a perfect student, but the turtles are perfect in terms of executing IRT-1PL. **Figure 5.18** is the logistic Q-Q plot for the mean of the changes in turtle θ values across years. Once again and not surprisingly, the fit is excellent which would indicate that there is both a ceiling limit and that the turtles are regressing to their mean.

The uncanny resemblance in the model longitudinal trends to those of real students indicates that the students' longitudinal performance on the TAKS exam is largely due to the IRT-1PL theoretical framework. Since the model represents a flawless execution of IRT-1PL, this means that the TAKS administrations are very close to being perfectly executed. There are many troublesome implications for this finding.

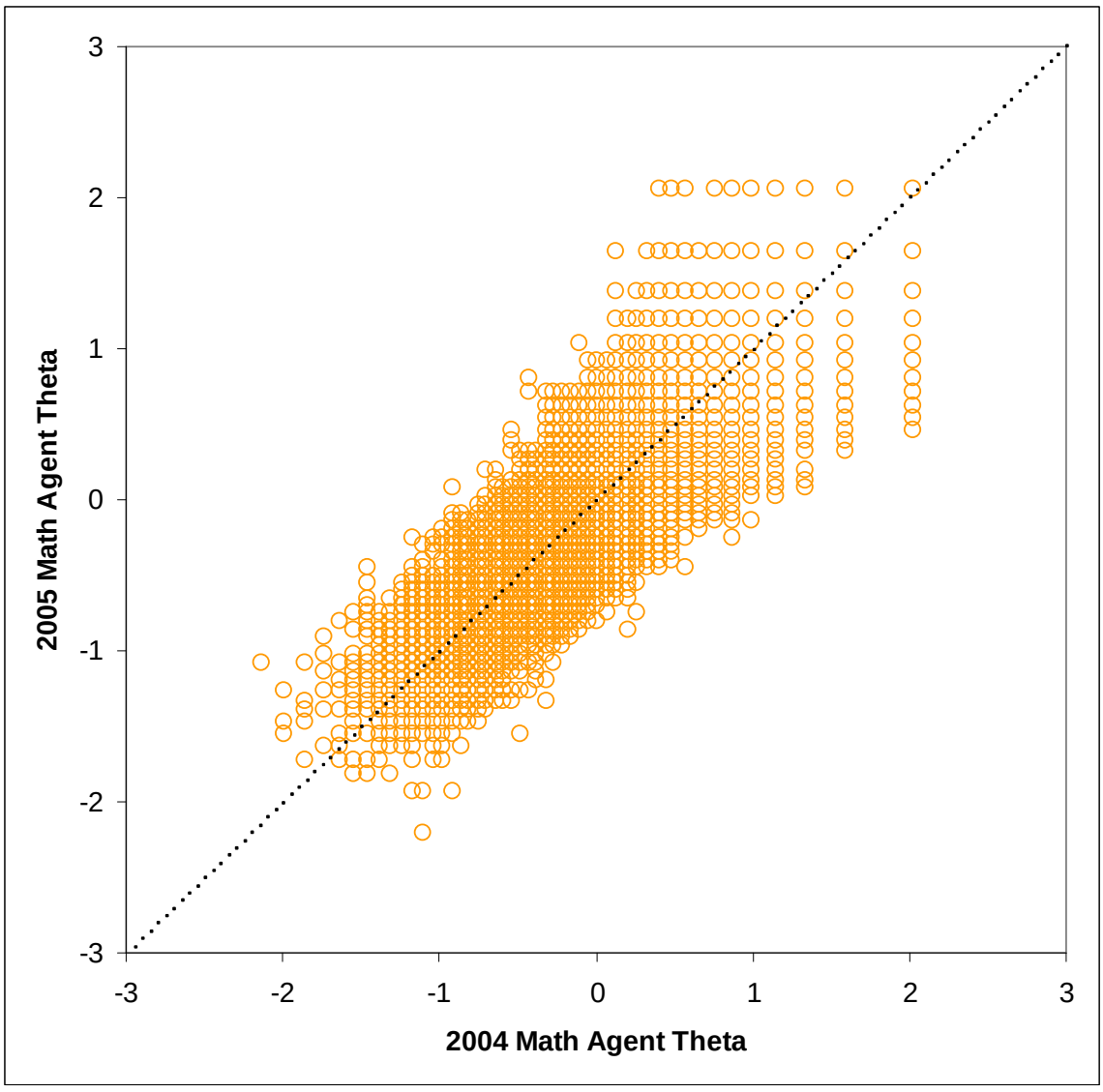

Figure 5.12 Graph of the 2005 Math Agent θ as a function of the 2004 Math Agent θ
(Fit Line: $y = 1.001x - 0.007$, $R^2 = 0.757$)

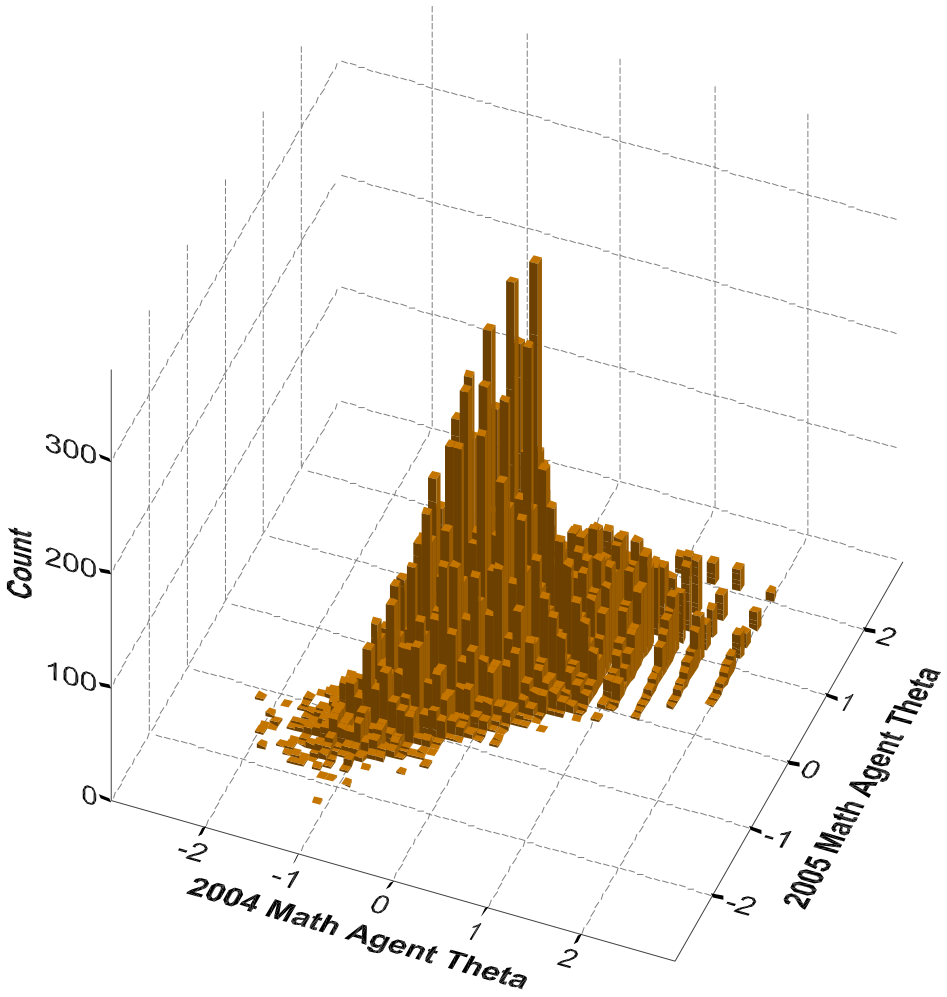

Figure 5.13 3D histogram of the 2005 Math Agent θ as a function of the 2004 Math Agent θ

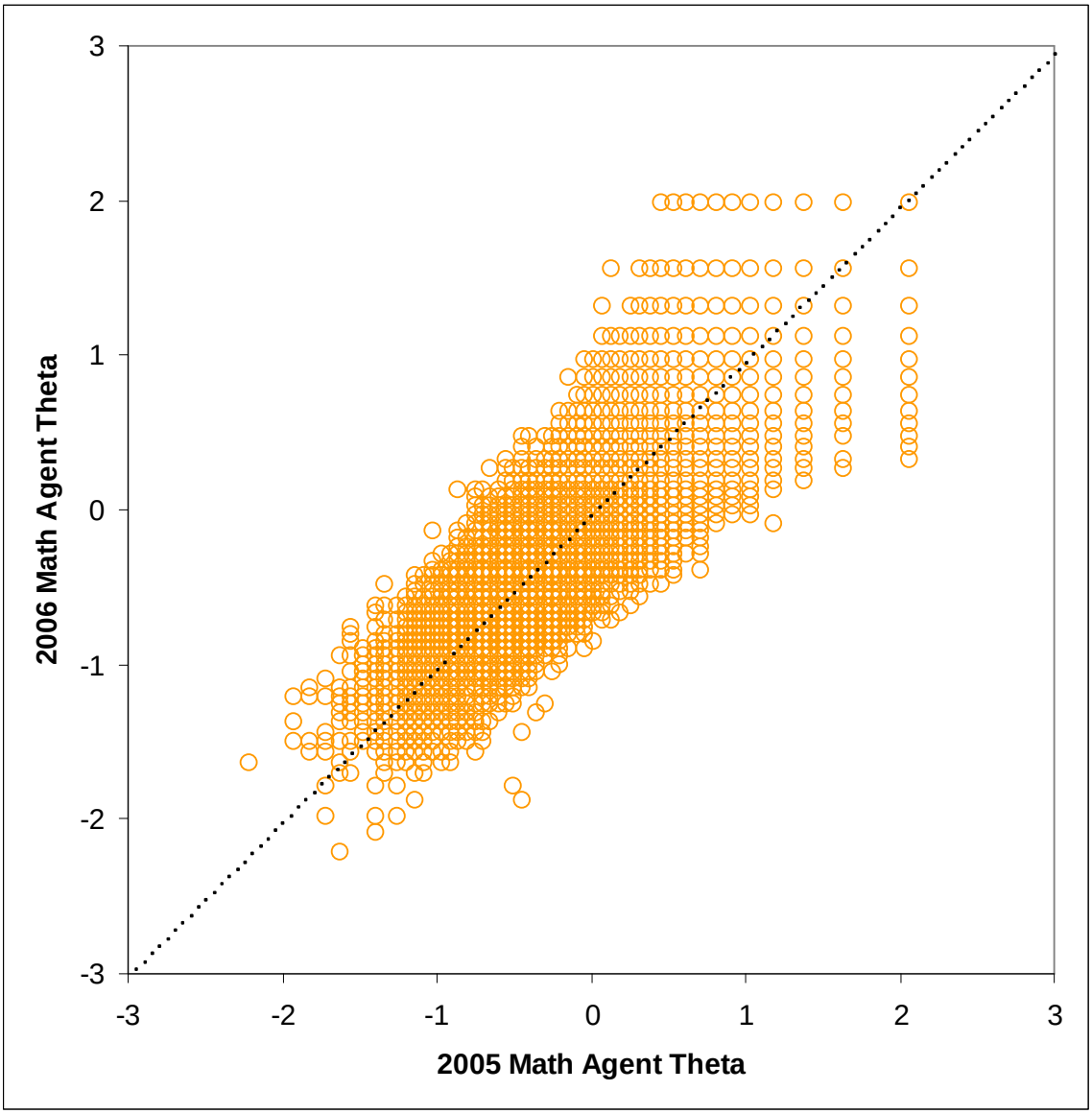

Figure 5.14 Graph of the 2006 Math Agent θ as a function of the 2005 Math Agent θ
(Fit Line: $y = 0.994x - 0.043$, $R^2 = 0.763$)

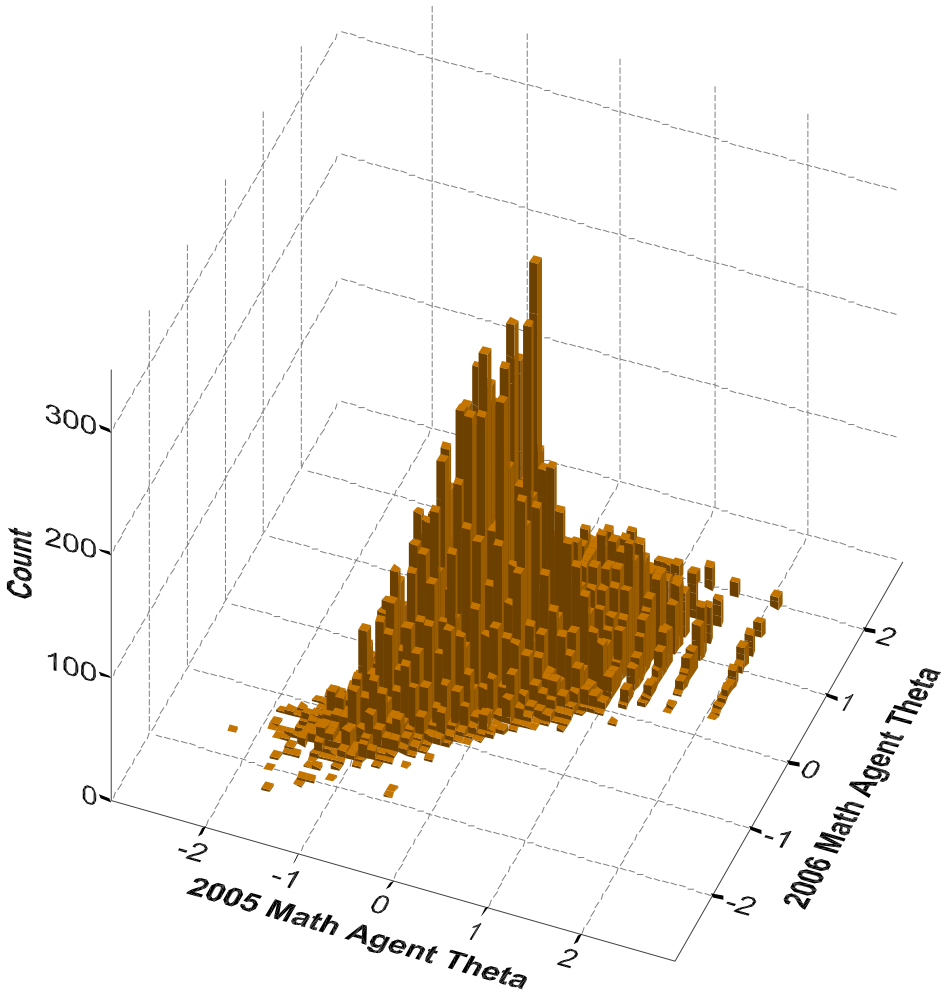

Figure 5.15 3D histogram of the 2006 Math Agent θ as a function of the 2005 Math Agent θ

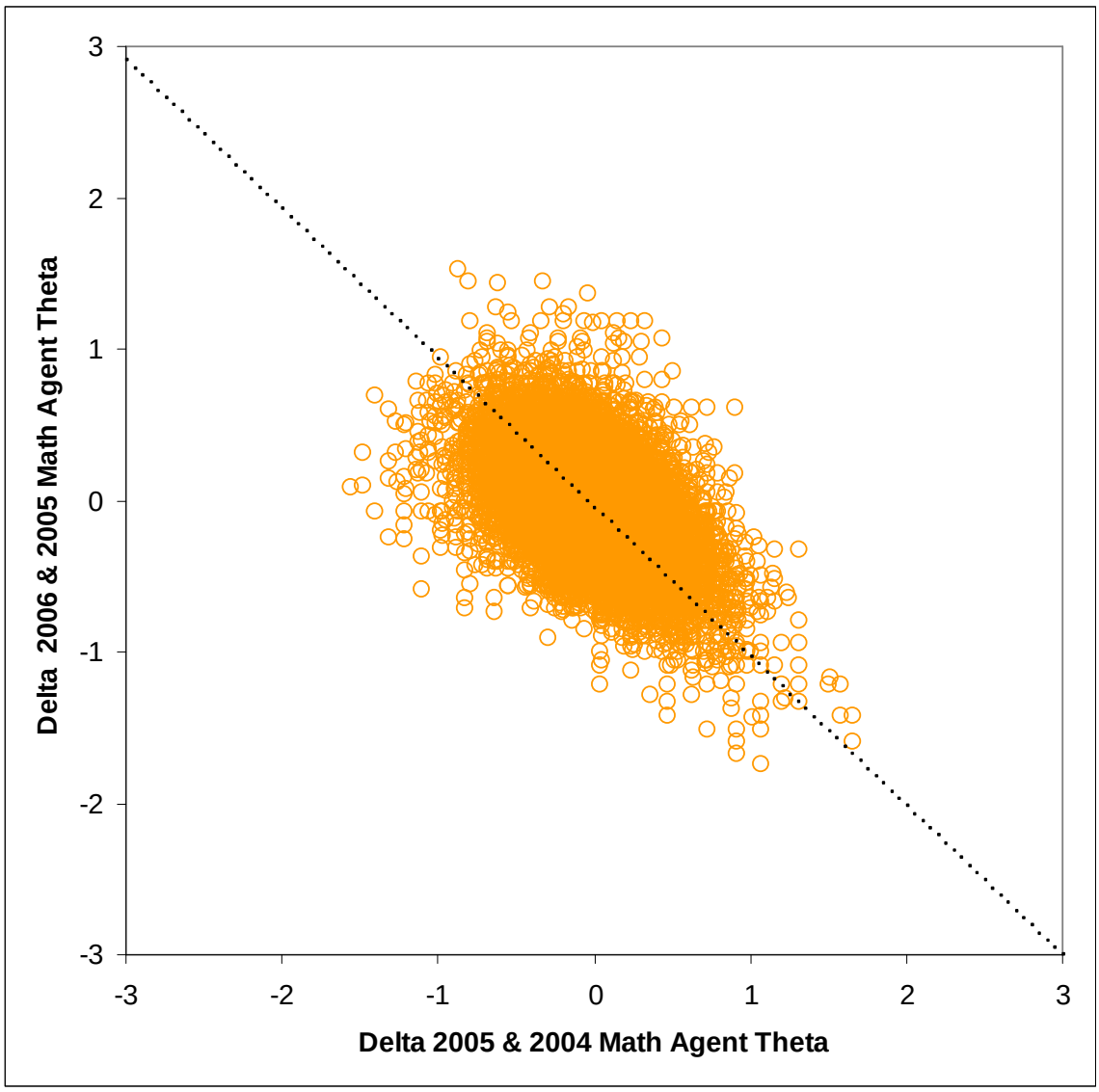

Figure 5.16 Graph of the difference between 2006 and 2005 Math Agent θ as a function of the difference between 2005 and 2004 Math Agent θ
(Fit Line: $y = -0.985x - 0.050$, $R^2 = 0.253$)

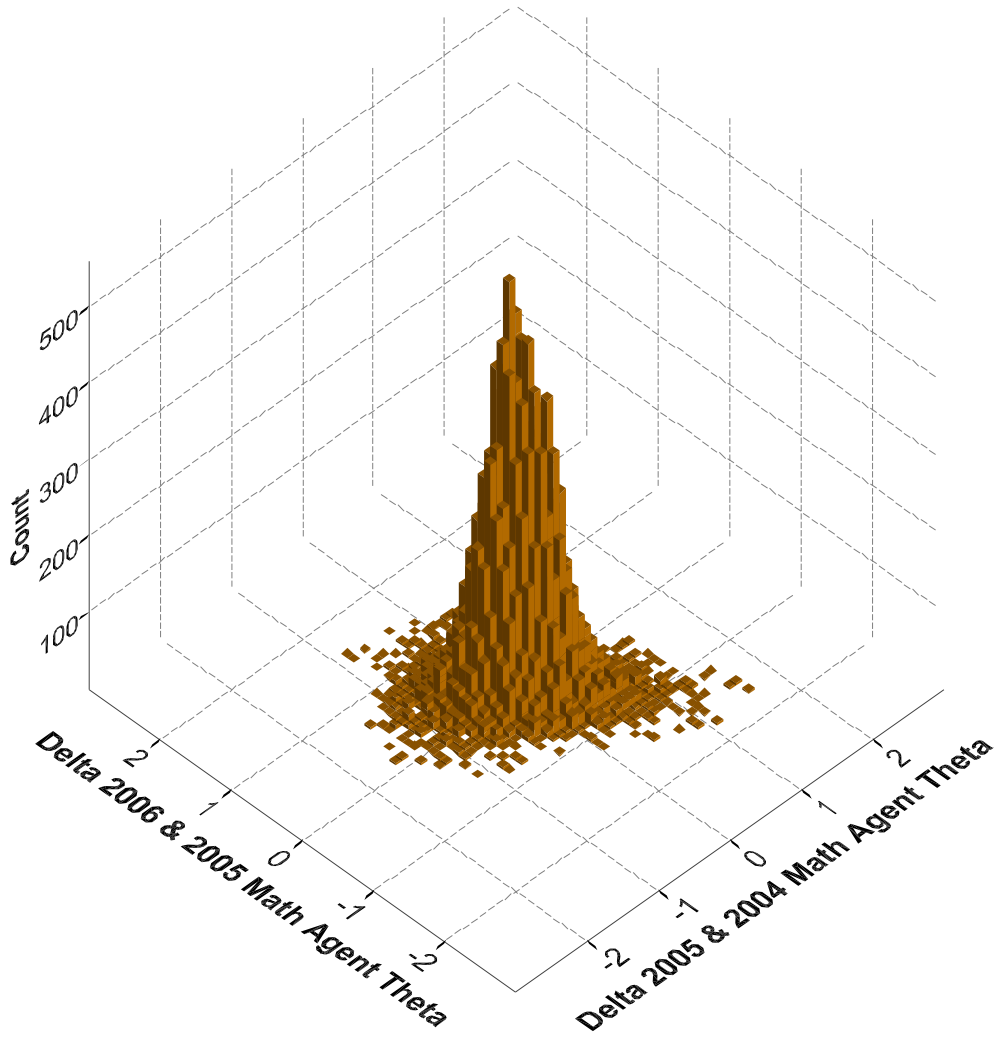

Figure 5.17 3D histogram of the difference between 2006 and 2005 Math Agent θ as a function of the difference between 2005 and 2004 Math Agent θ

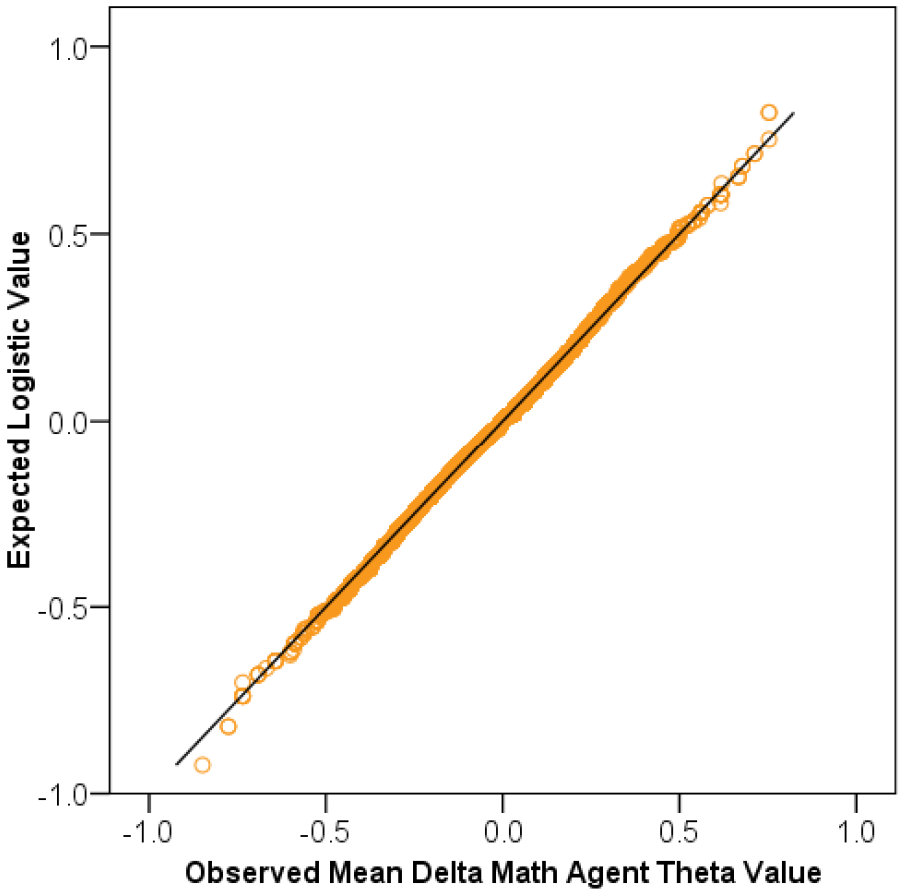

Figure 5.18 Logistic Q-Q plot of the mean change in math agent θ values

	R04 Agent θ	R05 Agent θ	R06 Agent θ	LTR	M04 Agent θ	M05 Agent θ	M06 Agent θ	LTM	H05 Agent θ	H06 Agent θ	LTH	S05 Agent θ	S06 Agent θ	LTS	LTP
R04 Agent θ	1.00														
R05 Agent θ	**0.70**	1.00													
R06 Agent θ	**0.68**	**0.69**	1.00												
LTR	0.83	0.84	0.82	1.00											
M04 Agent θ	0.56	0.57	0.55	0.68	1.00										
M05 Agent θ	0.57	0.57	0.56	0.68	**0.87**	1.00									
M06 Agent θ	0.57	0.58	0.56	0.69	**0.87**	**0.87**	1.00								
LTM	0.61	0.61	0.60	0.73	0.93	0.93	0.94	1.00							
H05 Agent θ	0.58	0.59	0.58	0.70	0.69	0.70	0.70	0.75	1.00						
H06 Agent θ	0.59	0.60	0.58	0.71	0.70	0.71	0.71	0.75	**0.83**	1.00					
LTH	0.65	0.66	0.64	0.78	0.77	0.77	0.77	0.83	0.91	0.91	1.00				
S05 Agent θ	0.63	0.64	0.62	0.76	0.75	0.75	0.76	0.81	0.78	0.79	0.86	1.00			
S06 Agent θ	0.63	0.64	0.62	0.76	0.75	0.76	0.76	0.81	0.78	0.79	0.86	**0.84**	1.00		
LTS	0.69	0.70	0.68	0.83	0.82	0.82	0.82	0.88	0.85	0.86	0.94	0.92	0.92	1.00	
LTP	0.69	0.70	0.68	0.83	0.82	0.82	0.82	0.88	0.85	0.86	0.94	0.92	0.92	1.00	1.00

Table 5.8 Correlation matrix of agent θ and latent trait values (N = 26,661)

Item Behavioral Trends

For item analysis, the data for the complete set of agents were used. **Figures 5.19** through **5.28** show the IRFs for the different domain items as a function of the different 2006 θ scales. Only the 2006 scales were used since the model assumes that the different domains are on the same scale and so the other years do not look any different. Using the Linkage values as determined from real students, the IRFs from the model are very similar to those seen in **Figures 4.12, 4.15, 4.17,** and **4.19**. This serves to justify both the use of the linear linkage **Equation 3.1** as well as the model used in the SEM of the real world from which the linkage equation was based. Even though the domain linkage values are not perfect, the IRFs indicate that there is enough shared commonalities across domains to yield items that behave in the same manner regardless of domain label. That does not mean that the items are not measuring some domain related content though, since there is some proportion of the latent traits that is unique to the domain in the model. In terms of the TAKS exam, it is this unique proportion that we should be measuring when we talk about achievement in the different domains.

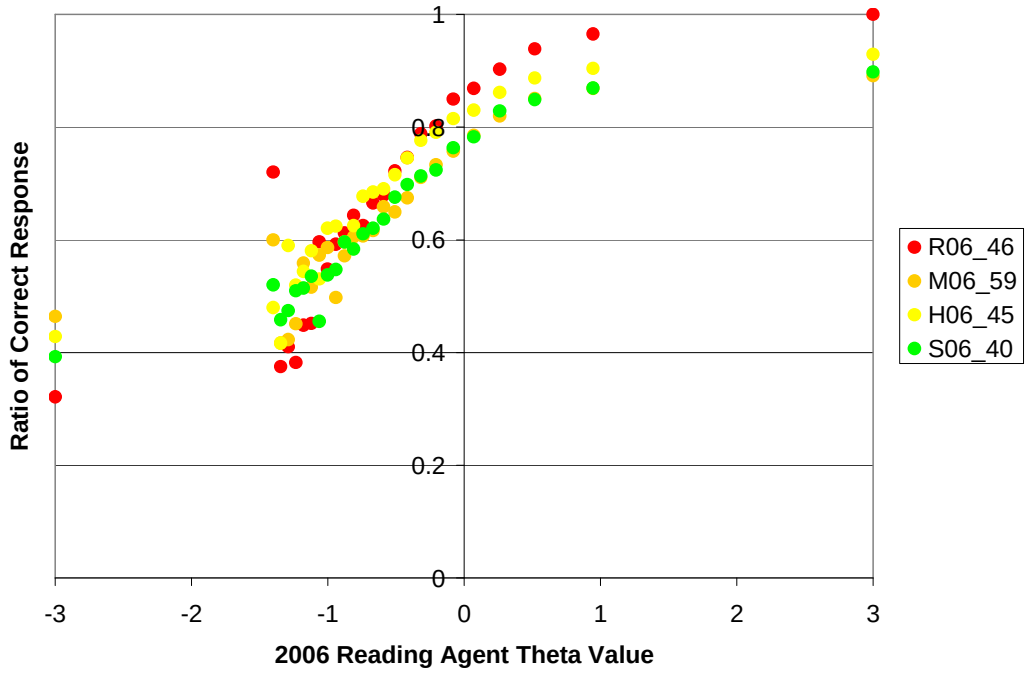

Figure 5.19 IRFs of the different domain items using R06 Agent θ for a simulated run

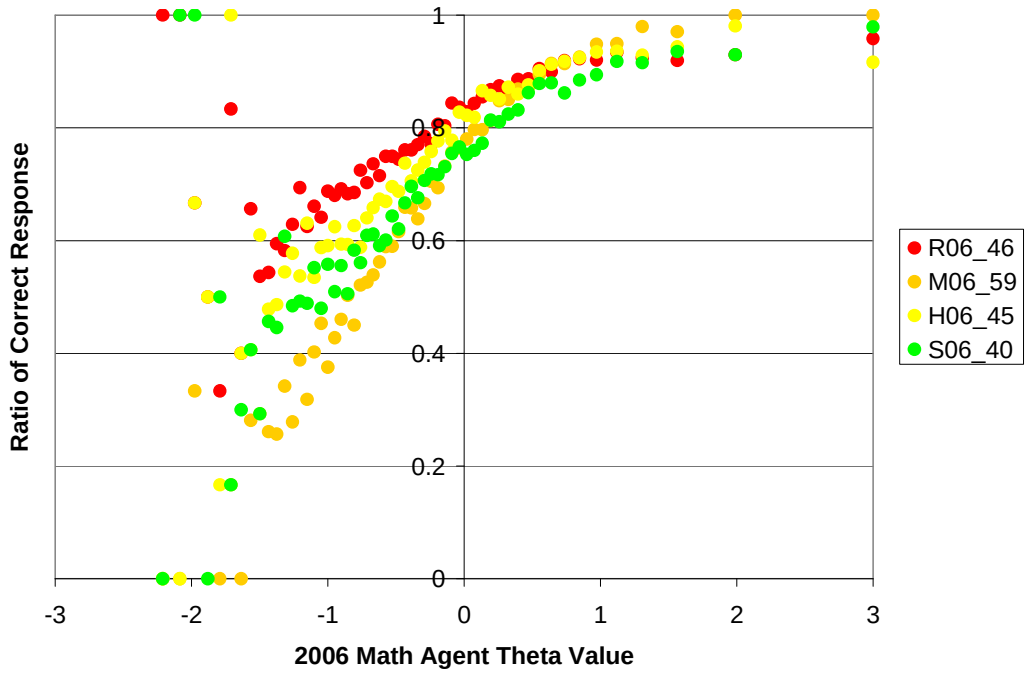

Figure 5.20 IRFs of the different domain items using M06 Agent θ for a simulated run

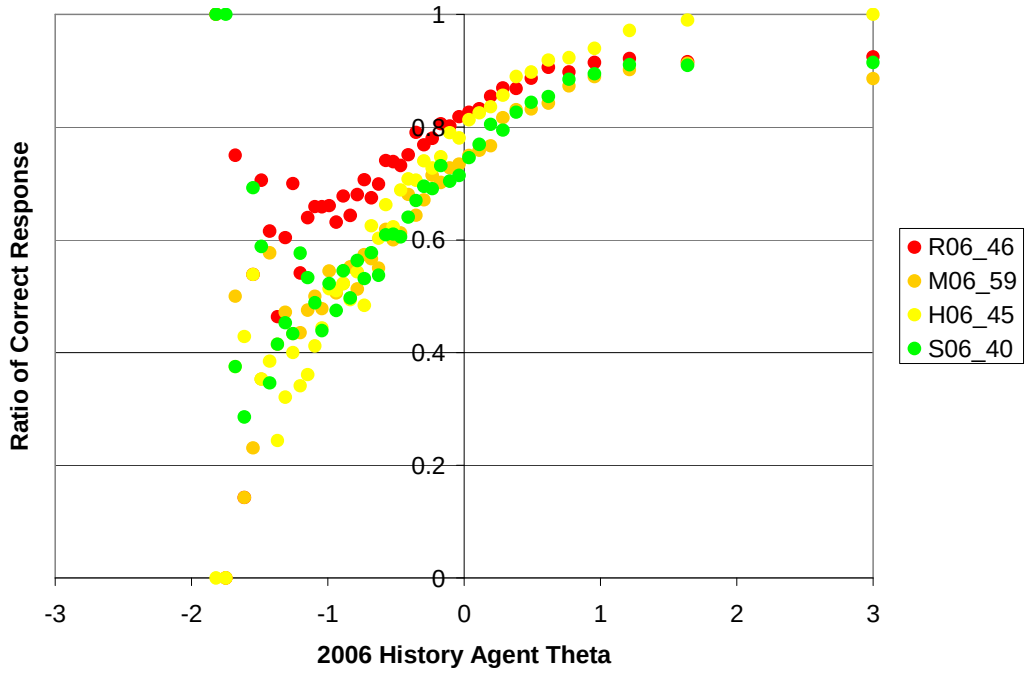

Figure 5.21 IRFs of the different domain items using H06 Agent θ for a simulated run

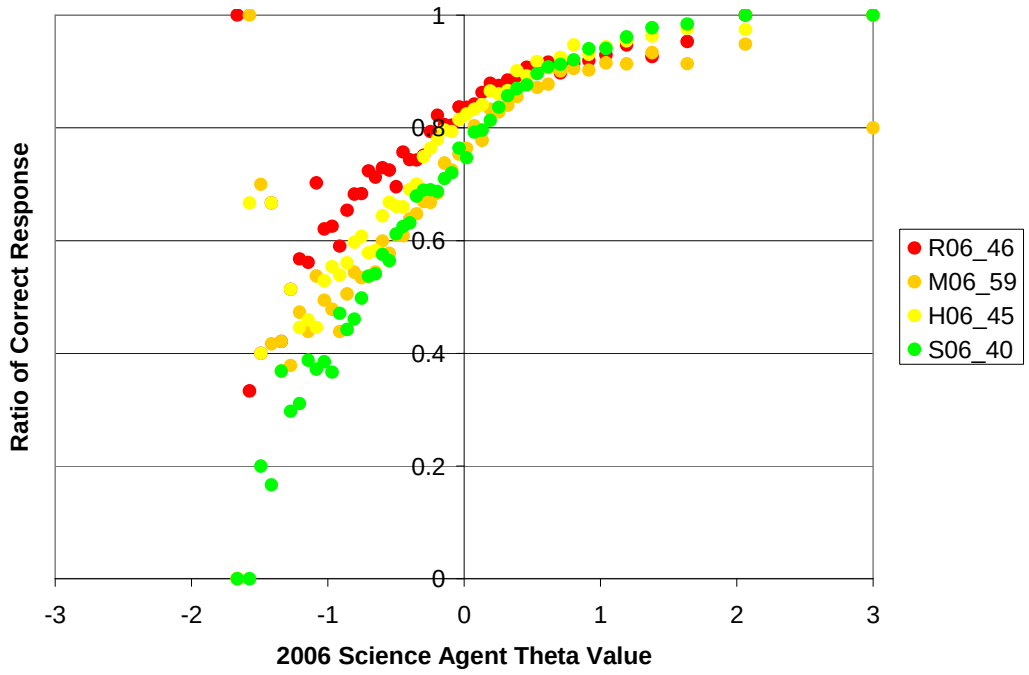

Figure 5.22 IRFs of the different domain items using S06 Agent θ for a simulated run

Students and TAKS Exam Interaction

To conclude the analysis of the simulation model data, we did an SEM analysis. The results can be seen in **Figure 5.23**. The first thing to notice is that unlike the real world data, this SEM shows a perfect fit to the data in every single measure of goodness-of-fit. One would expect this since the data is completely computer generated, based on the SEM of the real world data. This SEM is remarkably similar to the real one, both in terms of the amount of variance explained at each level as well as the correlation between variables. The "unexplained" variance of each section in the model data SEM is actually the error of measurement due to the IRT-1PL testing framework of the TAKS exam. This means that the difference between the real world data and the model data SEM represents the amount of variance that is truly unaccounted for by the real world data. This is also represents the amount of variance for which any teacher can hope to make a difference when it comes to improving TAKS scores. **Table 5.9** shows what these values are for each domain. Based on the fact that the trends on the TAKS exam and the model are very similar in nature, and assuming the model values are accurate, then very little can be done by the teachers to improve students' TAKS scores since the bulk of score is actually due to the IRT-1PL framework.

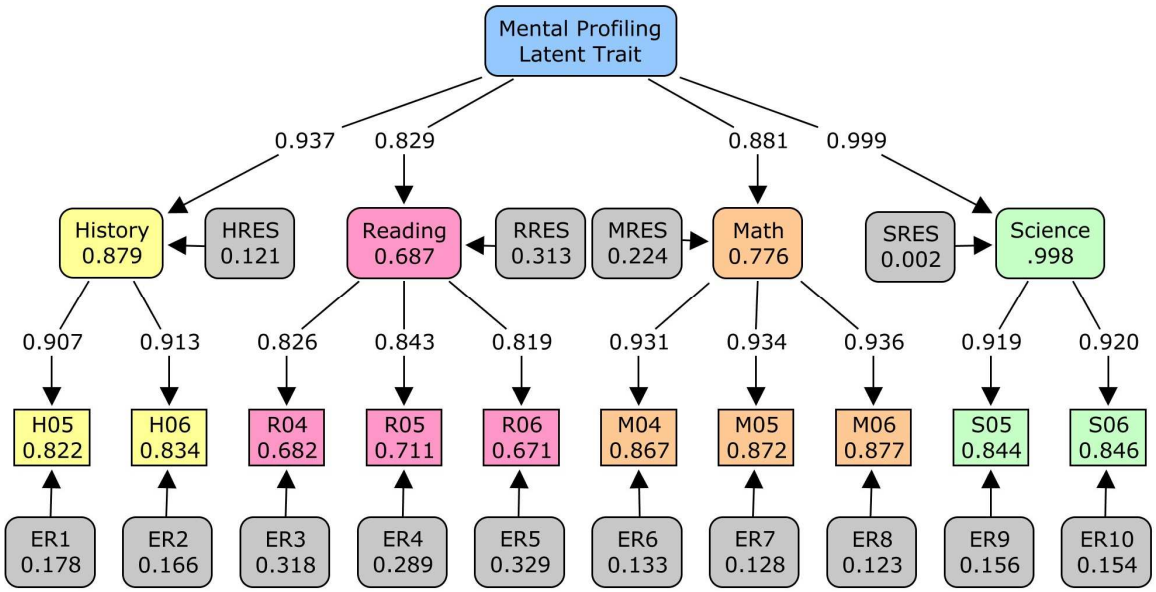

Figure 5.23 Structural Equation Modeling of Agent θ
(χ^2 = 32.337 (p = 0.401), N= 26,657, df = 31, RMSEA = 0.001, TLI = 1.000, NFI = 1.000, RFI = 1.000, IFI = 1.000, CFI = 1.000)

Domain and Year	Percent of Unaccounted Variance
R04	8.3
R05	6
R06	8
M04	8.8
M05	1.8
M06	11.1
H05	4.6
H06	14.6
S05	4.8
S06	9.6

Table 5.9 Percent of the variance on the TAKS exam that cannot be explained as random measurement error due to the IRT theoretical framework

Discussion

A model is only as good as the assumptions it is based on. In this dissertation, we have tried to faithfully reproduce the IRT-1PL model on a computer to see how it would behave in an ideal world. The codes for the computer model can be found in **Appendix B**. The codes are heavily commented to explain the purpose of each line of code. The bulk of the code is actually only concerned with setting up the initial agent population. The actual implementation of the IRT-1PL framework is relatively simple since the model is based on the probabilities of agents responding correctly to items based on their ability (θ value) and the difficulty (b-value) of the items as given in **Equation 2.4**. This means that there is only one source of error possible in the execution of the IRT model: the estimation of the b-values. The b-values of the items were determined based on the Analytic dataset, and are in perfect alignment with the scales used by the TAKS exam **(Table 4.5)**. Therefore, it is not expected that the b-values will be the source of any deviation from what a perfect execution of the 2004-2006 TAKS exams would look like.

The other variable in the IRT-1PL model is the θ value. The model allows for the user to flexibly set up the initial population as desired. In order to model the real world data, it is important to set up the agent population with the same distribution of θ values as to the Analytic dataset. Here, the assumption was made that each section within the domain actually tested for that domain, and so they can be summed together. The assumption makes sense if the sections are

truly testing for their domain label. The reason for this assumption is that the model only allows the user to set up the domain latent traits as a whole for each agent and not by each year of the TAKS exam. The assumption was justified when the gender data showed that the trends in each domain were different but consistent across years (**Figures 4.32** and **4.34**). The mean and standard deviation were determined from the summed scores in each domain and the turtle population set up to reproduce these values. Raw scores were used instead of θ values when determining the initial population values, since this did not require estimation of the turtle θ values from the response set to see if they meet the distribution from the Analytic dataset. Furthermore, by using raw scores which are on intervals of one item, the mean and standard deviations are more true to the population than the θ values when setting up the agent population. The reason is that a large number of students received a perfect score on one or more sections. These students did not get their θ values measured, because they are outside the range of the measurement scale but were given an arbitrarily high value to indicate that they performed at the maximal level. The mean and standard deviation of the internal θ values would not setup the correct distribution of θ values. Critics may point out summing the scores of the different sections of a domain this does not take into account the changes that students undergo between years. However, given the persistence of the longitudinal trends in the Analytic dataset, this was not a concern, especially given how well the model ended up mimicking the real world data.

In a perfect world, every exam would test for what it claims to be testing. We have seen what the IRFs of such tests would look like. Unfortunately, in the real world, it is not possible to create such perfect exams where the scores are only due to the domain tested. That does not mean we should not strive to create exams that maximizes on the content it is supposed to test. We have also shown what a poorly constructed exam would look like in terms of the IRFs. The TAKS exam is somewhere in between these two extremes. We have shown that the agent based modeling of the IRT-1PL framework of the TAKS exam on NetLogo yielded results that are very similar to real world data both in terms of the trends and values generated, which casts doubt on whether the TAKS exam really tests for achievement. If the TAKS exam allows little leeway for teachers to intervene, then the whole point of the No Child Left Behind Act becomes meaningless except as a way to penalize teachers who happen to have the wrong profile of students in their classrooms.

CHAPTER 6: Cross Validation and Model Confirmation

Recall that the original Longitudinal dataset was divided in half for the purposes of analysis and cross validation. Validation is important if we are to lend credibility to the model results. For one, it should be verified that the scales generated by the model and the real world data have not changed due to all of the intervening estimations. Furthermore, it needs to be proven that the trends seen in the real world data are a result of IRT-1PL. Even though the trend seen in the model closely mimics those seen in the real world data, this is not proof that those trends are due to IRT-1PL, though it is very likely. One way to prove this would be to determine how well the model is able to predict real student performance by comparing the predictions to the actual data in confirmatory analyses. This chapter will focus on validating the computer model so as to generate confidence in the conclusions drawn from the model data. It has even been suggested that the predicted data be used in a type of Turing test to see if relevant parties such as teachers and psychometricians can distinguish between model generated and real scores.

Validation of the Scales

One of the things that might have changed between all the estimations is the θ scales. When using the internal b-values and the response set from the model to estimate θ values, it should be determined if the scale changes at all. In order to do so, the Cross Validation dataset was subjected to the same procedures used on the model Simulation dataset. The first step was to use the predetermined internal b-values from the Analytic dataset and the response set from the Cross Validation dataset to determine the student θ values with PARAM-1PL as was done with the model data. **Table 6.1** shows the results of the correlation matrix when student with perfect scores were removed. This table is virtually identical to **Table 4.1** with the estimated θ and scales scores perfectly correlated for all domains except for Reading. As such, using the internal b-values in the model does not change the ratio of the scales for student or turtle θ values from that of the TAKS scales. Furthermore, it is now also possible to convert agent θ values to scales scores if so desired via a linear transformation. The conversion equation is as follows:

$$\text{Scale Score} = \mathbf{M} * \text{Est. } \theta + \mathbf{B}$$

Equation 6.1 Conversion of PARAM-1PL estimated θ values to TAKS scale scores

M and **B** are constants and can be derived by doing linear regression on the Cross Validation students' scale scores and estimated θ values. These constants are listed in **Table 6.2** by domain and year. The Reading domain was omitted

since the essay items cannot be modeled and so a perfect conversion to scale scores from the estimated θ values is impossible.

A set of analyses using the estimated θ values, such as those done for the Analytic and the model Simulation data, was also done on the Cross Validation dataset. Although not presented here in the dissertation, suffice it to say that the results are identical to the real world data analyses since the ratio of the scales have not been altered by the use of internal b-values to estimate θ values in PARAM-1PL.

	R04 θ Est.	R04 SSC	R05 θ Est.	R05 SSC	R06 θ Est.	R06 SSC	M04 θ Est.	M04 SSC	M05 θ Est.	M05 SSC	M06 θ Est.	M06 SSC	H05 θ Est.	H05 SSC	H06 θ Est.	H06 SSC	S05 θ Est.	S05 SSC	S06 θ Est.	S06 SSC
R04 θ Est.	1.00																			
R04 SSC	**0.97**	1.00																		
R05 θ Est.	0.62	0.63	1.00																	
R05 SSC	0.55	0.57	**0.72**	1.00																
R06 θ Est.	0.59	0.60	0.61	0.52	1.00															
R06 SSC	0.55	0.57	0.56	0.58	**0.78**	1.00														
M04 θ Est.	0.55	0.57	0.56	0.50	0.54	0.50	1.00													
M04 SSC	0.55	0.57	0.56	0.50	0.53	0.50	**1.00**	1.00												
M05 θ Est.	0.54	0.56	0.57	0.52	0.55	0.52	0.82	0.82	1.00											
M05 SSC	0.54	0.56	0.57	0.52	0.55	0.51	0.82	0.82	**1.00**	1.00										
M06 θ Est.	0.50	0.51	0.52	0.47	0.53	0.50	0.77	0.77	0.81	0.81	1.00									
M06 SSC	0.50	0.51	0.52	0.47	0.53	0.50	0.77	0.77	0.81	0.81	**1.00**	1.00								
H05 θ Est.	0.61	0.62	0.63	0.54	0.57	0.53	0.64	0.64	0.68	0.68	0.61	0.61	1.00							
H05 SSC	0.61	0.62	0.63	0.54	0.57	0.53	0.64	0.64	0.68	0.68	0.61	0.61	**1.00**	1.00						
H06 θ Est.	0.56	0.56	0.56	0.47	0.53	0.48	0.58	0.58	0.60	0.60	0.60	0.60	0.73	0.73	1.00					
H06 SSC	0.56	0.56	0.56	0.47	0.53	0.48	0.58	0.58	0.60	0.60	0.60	0.60	0.73	0.73	**1.00**	1.00				
S05 θ Est.	0.57	0.58	0.60	0.52	0.55	0.51	0.72	0.72	0.76	0.76	0.70	0.70	0.75	0.75	0.68	0.68	1.00			
S05 SSC	0.57	0.58	0.60	0.52	0.55	0.51	0.72	0.72	0.76	0.75	0.70	0.70	0.75	0.75	0.68	0.68	**1.00**	1.00		
S06 θ Est.	0.54	0.55	0.56	0.48	0.54	0.49	0.68	0.68	0.71	0.71	0.73	0.73	0.70	0.70	0.73	0.73	0.77	0.77	1.00	
S06 SSC	0.54	0.54	0.56	0.48	0.54	0.49	0.68	0.68	0.71	0.71	0.73	0.73	0.70	0.70	0.73	0.73	0.77	0.77	**1.00**	1.00

Table 6.1 Correlation matrix for student actual scale scores and their θ values estimated by PARAM-1PL using internal b-values of the Analytic dataset (N= 60,606)

Domain & Year	M	B	R^2
M04	306.145	2160.840	1.000
M05	235.327	2149.255	1.000
M06	236.128	2208.002	1.000
H05	243.715	2261.871	1.000
H06	213.565	2289.021	1.000
S05	268.987	2130.540	1.000
S06	217.616	2183.436	1.000

Table 6.2 Constant to transform estimated θ values to TAKS scale scores

Using Real World Distributions in a Model World

Another way in which the model was validated was to use real world values and distributions in the model. To do this, a random subset of 30,000 students was chosen from the Cross Validation dataset. An estimate of their domain latent traits had to be determined. This was done in one of two ways. The first is to take the average of all of their estimated θ values within each domain. In theory, the score on each section within a domain represents an estimate of the magnitude of that domain latent trait as well as incorporating some amount of measurement error. Taking the average would give a value that would be closer to the true value by reducing the amount of measurement error, assuming that it is random. Alternatively, the response set of all items within a domain could be used to estimate the domain latent trait value. Due to the larger number of items used to measure the student on each domain latent trait scale across years, the values derived should have very little errors. This latter option was chosen since it represented one accurate measure, as opposed to being the average of several less accurate measures. Regardless of which option was chosen, the values are almost perfectly correlated, as shown in **Table 6.3**. Once the latent trait values were determined, they were then converted to the scale used in the simulation model run (**Table 5.5**). This is so that the scales in the model match the scales used by the TAKS exam in terms of generating the desired distribution output. The model can now be run using the real students as if they were turtles in the model.

This type of validation is useful because it allows for a direct comparison between model and real world results. Both model and real world data are a result of the IRT-1PL theoretical framework. If the output from the model closely resembles the real data for the students, it would indicate that the computer model operates similarly to the principles governing the behavior of students on the TAKS exam in the real world. A congruence of results between the two would mean that TAKS scores are mostly due what is in common: the IRT-1PL framework. Furthermore, if we assume that the model is the representation of a perfect execution of an IRT-1PL exam, then it would provide a baseline comparison to how well IRT-1PL is being implemented by the TAKS exam. There is one caveat though and that is that the "actual" value of LTP as well as the linkage values to each domain will be unknown whereas it is normally declared under a normal model run. Both of these are greatly affected by errors of estimation in the domain latent traits. Any error component in the estimated domain latent trait values will show up as unexplained variance in that domain on the SEM analysis.

	LTR A	LTR B	LTM A	LTM B	LTH A	LTH B	LTS A	LTS B
LTR A	1.00	0.97	0.72	0.72	0.75	0.74	0.74	0.73
LTR B	**0.97**	1.00	0.70	0.70	0.72	0.72	0.71	0.71
LTM A	0.72	0.70	1.00	1.00	0.75	0.75	0.84	0.84
LTM B	0.72	0.70	**1.00**	1.00	0.75	0.75	0.84	0.84
LTH A	0.75	0.72	0.75	0.75	1.00	0.99	0.84	0.84
LTH B	0.74	0.72	0.75	0.75	**0.99**	1.00	0.83	0.83
LTS A	0.74	0.71	0.84	0.84	0.84	0.83	1.00	1.00
LTS B	0.73	0.71	0.84	0.84	0.84	0.83	**1.00**	1.00

Table 6.3 Correlation matrix of the estimated domain latent traits A = PARAM-1PL estimated from entire set of all domain items, B = average θ value for each domain

Student Behavioral Trends

One way in which we can see how well the model predicts the real world is to see how the well estimated θ and model θ values correlate with each other as well as to the domain latent trait estimates. **Table 6.4** shows the correlation matrix of the real and model θ values. The correlations between the estimated domain latent traits and the real θ values are higher than to the model derived θ values. Since the domain latent trait values were estimated from the total domain response set, this is not surprising. The fact that the model θ values are not as well correlated is a result of being the estimate of an estimate. Note, however, that the model θ values are better correlated with each other within a domain than the real θ values do since they are due to a single underlying latent trait as is assumed in the model. This means that there are variances in between the sections in each domain that is not explained by the domain latent trait. Recall that these values were shown in **Table 5.9** based on the model Simulation data. While the assumption made in the computer model was that the domain latent traits are solely responsible for the scores on each section, it is apparent there and here that there are other sources of variation in the scores for each section even if they are small.

Figure 6.1 shows the 2005 Math θ values as a function of the 2004 Math θ values for both the real and model data. We have already seen that the model can generate fit lines and R^2 values similar to those in the real world from **Chapter 4.** Here we have a direct comparison and it is undeniable that the model is very good at simulating how students score from year to year with students

maintaining their relative position on the scale of measurement. **Figure 6.2** graphs the differences between each year and it can be seen that the model simulates the changes across years to real world values very well also. Item analysis was not done since it adds nothing more to the current discussion.

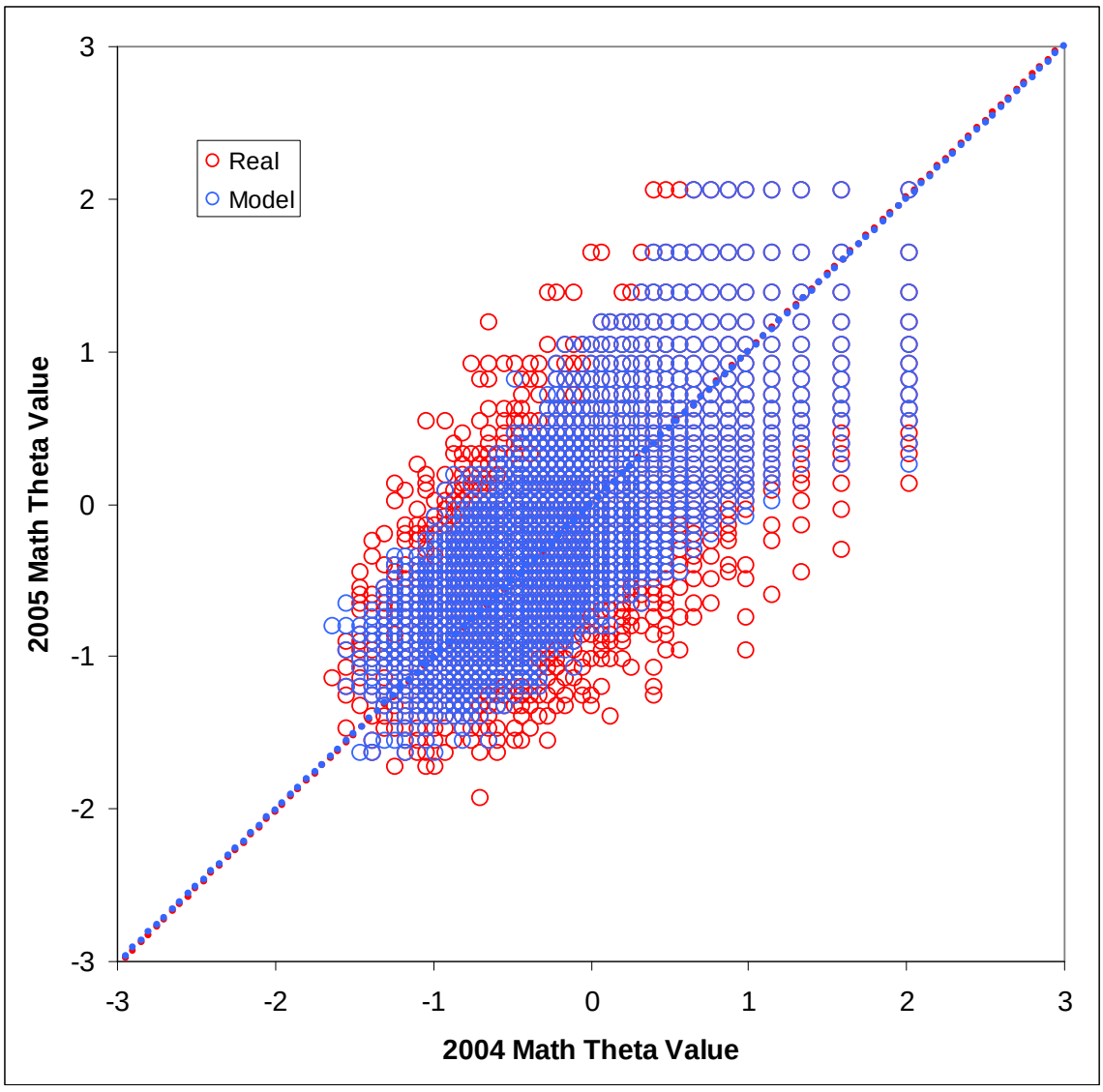

Figure 6.1 Graph of the 2005 θ value as a function of the 2004 θ value for both real and model data
(Real Fit Line: y = 1.007x - 0.006, R^2 = 0.675
Model Fit Line: y = 1.002x - 0.005. R^2 = 0.732)

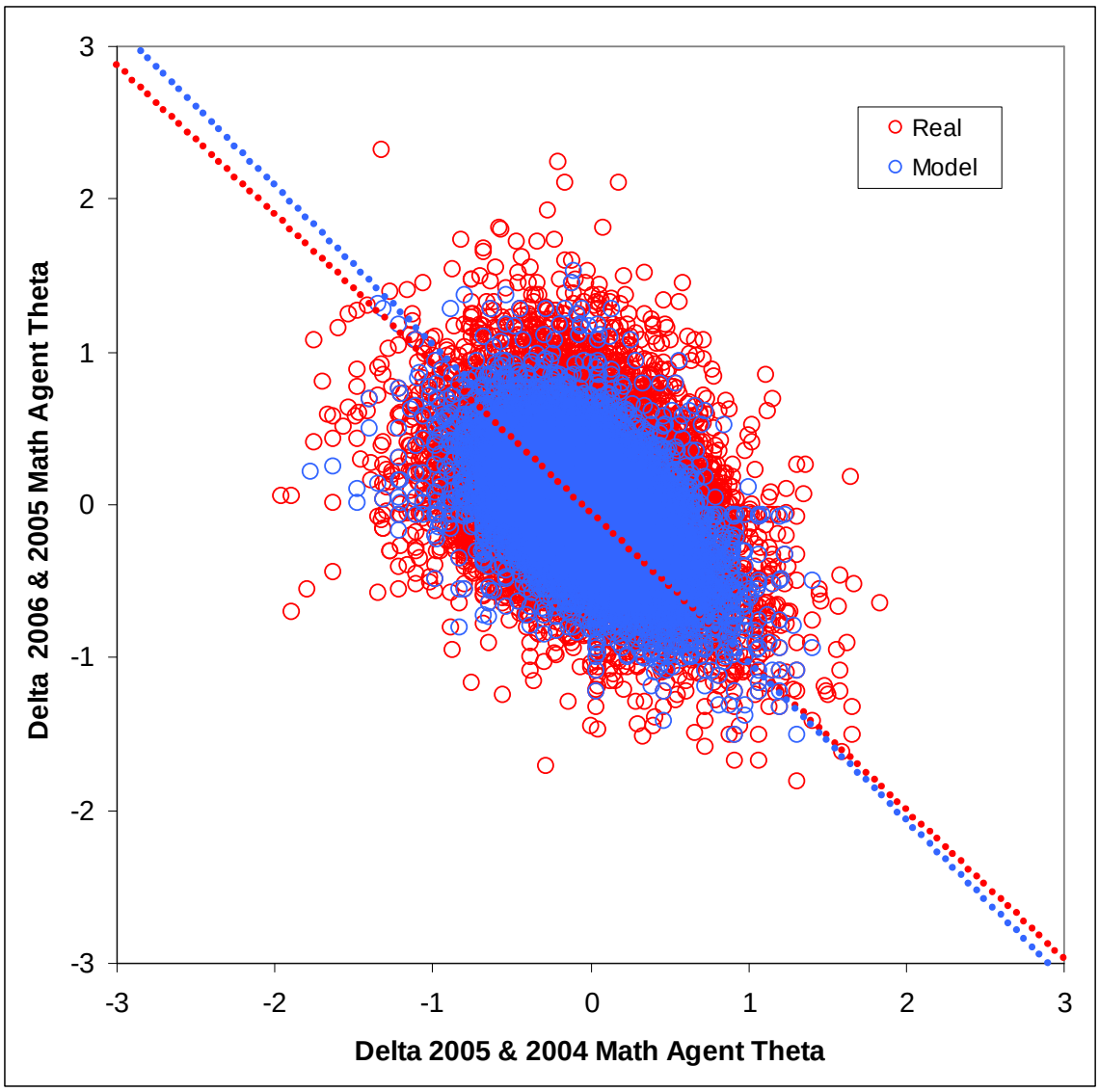

Figure 6.3 Graph of the difference between the 2006 and 2005 θ values as a function of the difference between the 2005 and 2004 θ values
(Real Fit Line: $y = -1.038x + 0.013$, $R^2 = 0.153$
Model Line: $y = -0.974x - 0.047$, $R^2 = 0.254$)

	R04 M θ	R04 R θ	R05 M θ	R05 R θ	R06 M θ	R06 R θ	LTR	M04 M θ	M04 R θ	M05 M θ	M05 R θ	M06 M θ	M06 R θ	LTM	H05 M θ	H05 R θ	H06 M θ	H06 R θ	LTH	S05 M θ	S05 R θ	S06 M θ	S06 R θ	LTS
R04 M θ	1.00																							
R04 R θ	**0.68**	1.00																						
R05 M θ	0.66	0.70	1.00																					
R05 R θ	0.69	0.59	**0.71**	1.00																				
R06 M θ	0.64	0.68	0.65	0.68	1.00																			
R06 R θ	0.65	0.56	0.67	0.58	**0.65**	1.00																		
LTR	0.80	0.86	0.82	0.86	0.80	0.82	1.00																	
M04 M θ	0.48	0.51	0.50	0.53	0.49	0.51	0.61	1.00																
M04 R θ	0.50	0.53	0.52	0.54	0.50	0.52	0.63	**0.84**	1.00															
M05 M θ	0.49	0.51	0.51	0.53	0.49	0.52	0.61	0.85	0.85	1.00														
M05 R θ	0.50	0.52	0.52	0.55	0.50	0.53	0.63	0.86	0.81	**0.87**	1.00													
M06 M θ	0.49	0.51	0.51	0.53	0.49	0.52	0.62	0.85	0.85	0.85	0.87	1.00												
M06 R θ	0.47	0.48	0.48	0.50	0.47	0.51	0.59	0.84	0.76	0.85	0.80	**0.85**	1.00											
LTM	0.53	0.55	0.55	0.58	0.53	0.56	0.67	0.92	0.92	0.92	0.94	0.92	0.92	1.00										
H05 M θ	0.51	0.54	0.52	0.56	0.50	0.51	0.63	0.57	0.56	0.57	0.59	0.57	0.56	0.62	1.00									
H05 R θ	0.55	0.59	0.57	0.61	0.55	0.55	0.69	0.62	0.62	0.62	0.66	0.63	0.59	0.68	**0.82**	1.00								
H06 M θ	0.51	0.54	0.52	0.56	0.50	0.52	0.64	0.57	0.57	0.57	0.60	0.57	0.56	0.62	0.79	0.82	1.00							
H06 R θ	0.50	0.53	0.52	0.54	0.49	0.51	0.63	0.56	0.56	0.57	0.58	0.57	0.58	0.62	0.82	0.71	**0.82**	1.00						
LTH	0.57	0.61	0.59	0.63	0.56	0.58	0.71	0.64	0.64	0.64	0.67	0.65	0.63	0.70	0.89	0.92	0.89	0.93	1.00					
S05 M θ	0.49	0.51	0.52	0.55	0.50	0.51	0.62	0.67	0.66	0.67	0.69	0.67	0.67	0.73	0.65	0.69	0.65	0.67	0.73	1.00				
S05 R θ	0.52	0.55	0.55	0.59	0.53	0.54	0.66	0.70	0.70	0.71	0.74	0.71	0.68	0.77	0.67	0.74	0.68	0.66	0.76	**0.85**	1.00			
S06 M θ	0.49	0.51	0.51	0.55	0.49	0.51	0.62	0.67	0.66	0.67	0.69	0.67	0.68	0.73	0.65	0.69	0.65	0.67	0.73	0.82	0.85	1.00		
S06 R θ	0.49	0.51	0.51	0.55	0.50	0.52	0.62	0.68	0.66	0.68	0.69	0.69	0.71	0.74	0.67	0.68	0.67	0.72	0.76	0.85	0.76	**0.85**	1.00	
LTS	0.54	0.57	0.57	0.61	0.55	0.56	0.68	0.74	0.73	0.74	0.76	0.74	0.75	0.81	0.72	0.76	0.72	0.74	0.81	0.90	0.94	0.90	0.94	1.00

Table 6.4 Correlation matrix of the estimated real world and the model θ values (N = 24,575) R = Real and M = Model

Student and TAKS Exam Interaction

Perhaps the most interesting analysis to compare is the SEM between the real world and model data since it shows how the variances are shared. **Figure 6.3** shows the SEM of the real world data for the selected students. Note how similar it is to **Figure 4.24**. Since both figures use real world data, this is not surprising at all. **Figure 6.4** shows the SEM of the model data. Note how similar the amounts of variance that can be explained in each section are with those in **Figure 6.3**. Just like in **Figure 5.23**, the SEM can explain more variance in each section on the computer model than in real life. This is due to the assumption that one domain latent trait explains all of the variance in each section of that domain causing an increase in the intra-domain correlations in the model. The increased intra-domain correlation means that more variance of each section gets explained by the domain latent trait. Also note how the amounts of variance that can be explained in the domain latent traits have decreased in the model SEM. This is the result of using an estimated domain latent trait value in the model. The reduced correlation of the model θ value to domain latent trait value as compared to the real θ values in **Table 6.4** is a result of estimation causing the shared variance across domain latent traits to be reduced. This is one reason why SEM is so powerful. The estimation of latent traits by looking at only shared variances means that there is very little error of measurement in the latent traits. In this case, the real world data can explain more of the variance than the model data for the domain latent traits. The domain latent traits in the real world SEM have practically no measurement error whereas the model SEM had to be estimated

with some level of error to be used in the computer model, then run in the model, and finally re-estimated in the SEM. It is not surprising that this would yield lower correlations with LTP for the domain latent traits. Still, when one considers the difference between the amounts of variance that is accounted for in the model as compared to the real world, the error of estimation cannot be very much.

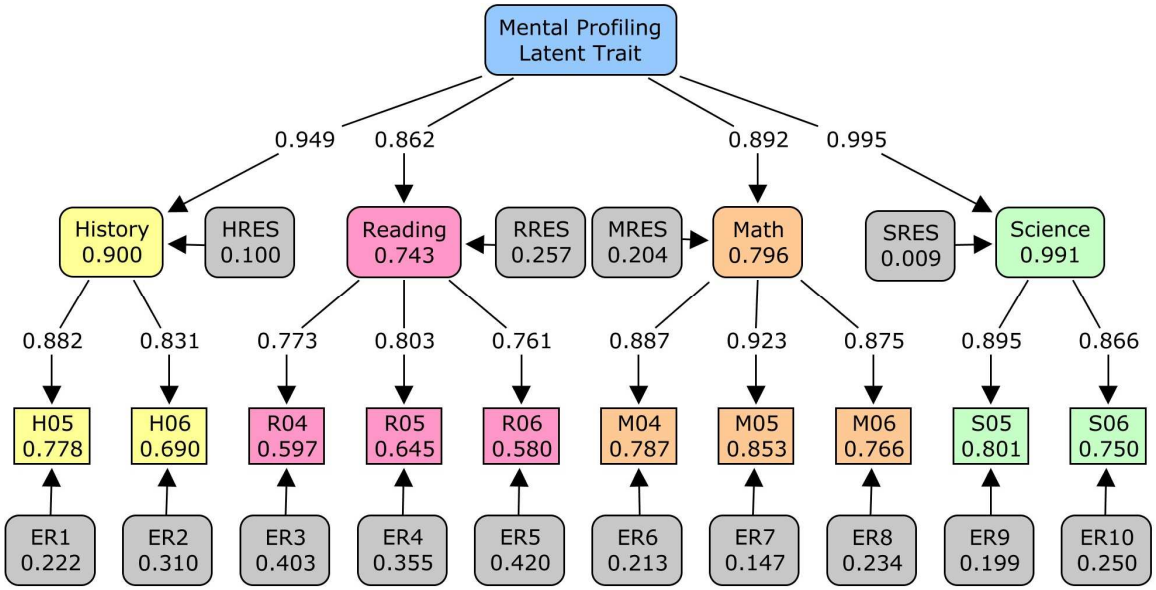

Figure 6.3 Structural Equation Modeling for the Real world data of the Cross Validation students used in the computer model (χ^2 = 5,059.195 ($p < 0.001$), N= 26,188, df = 31, RMSEA = 0.079, TLI = 0.966, NFI = 0.977, RFI = 0.966, IFI = 0.977, CFI = 0.977)

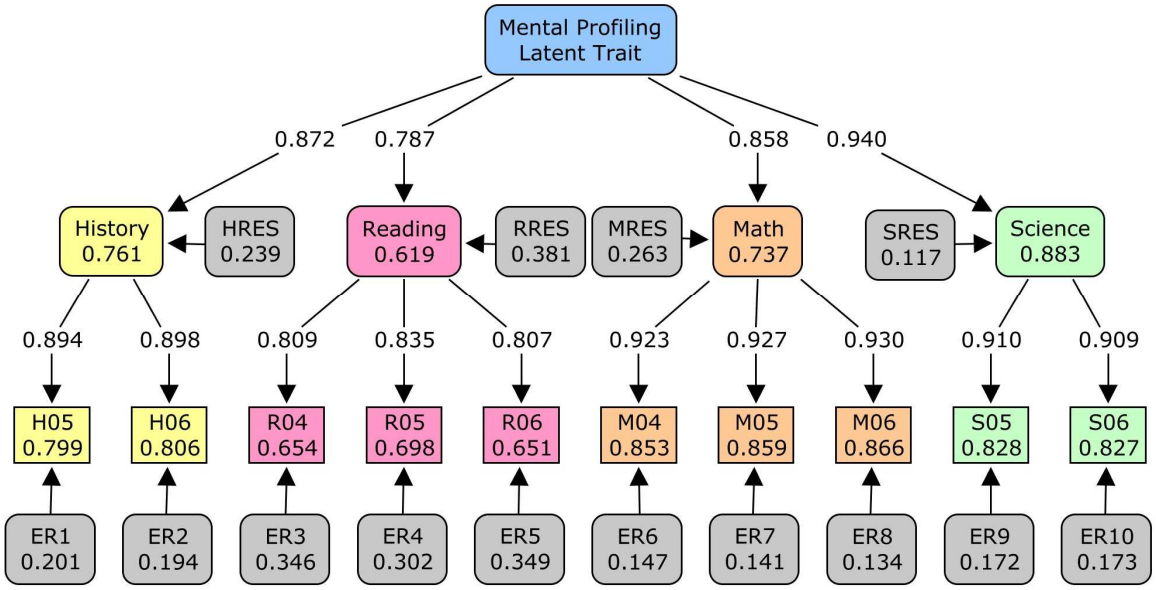

Figure 6.4 Structural Equation Modeling for the Model data of the Cross Validation students used in the computer model (χ^2 = 955.791 ($p < 0.001$), N= 26,351, df = 31, RMSEA = 0.0034, TLI = 0.994, NFI = 0.996, RFI = 0.994, IFI = 0.996, CFI = 0.996)

Comparison of Real World and Model Data

We have shown in **Chapter 4** and **5** that the mean change in scores from year to year is logistically distributed indicating both RTM and a ceiling effect. If we now assume that the expected scores from the model represent the perfect execution of the IRT-1PL framework, then the difference between the students' model scores and their real world scores would indicate how well the IRT-1PL framework was implemented in the real world. **Figure 6.5** is the graph of the real θ values as a function of the model θ values. The slope of the fit line is close to unity and can explain over 70% of the variance in the data. Considering that in the pure model results, only ~24% of the variance is unexplained (**Figures 5.12** and **5.14**) and is due entirely to random measurement error, that leaves about 4% of the error that is truly unexplained between the expected and observed θ values for these selected students. **Figure 6.6** shows the difference in the 2005 model and real θ values as a function of the 2004 Math model and real θ values. The fit line can only explain about 2% of the variance. This means that the differences between model and real θ values across years are also random. Finally, **Figure 6.7** is the logistic Q-Q plot of the mean difference between the Math model and real θ values for all years. Rather than fitting perfectly, the plot indicates that there are fewer students than would be expected in a logistic distribution whose difference between mean model and real θ values are large. Oddly, the trend is not consistent across the different domains. Reading seems to be normally distributed while History and Science, like Math, fit neither

distribution but rather the low end fits a logistic distribution while the high end fits to a normal distribution better. This indicates skewness in the distribution rather than the expected logistic distribution. The specific cause of the differences in terms of deviation from the IRT-1PL expected distribution is unknown. The differences between model and real θ values are not completely random. It can be concluded then that there is a slight deviation from a perfect IRT-1PL execution in the real world data.

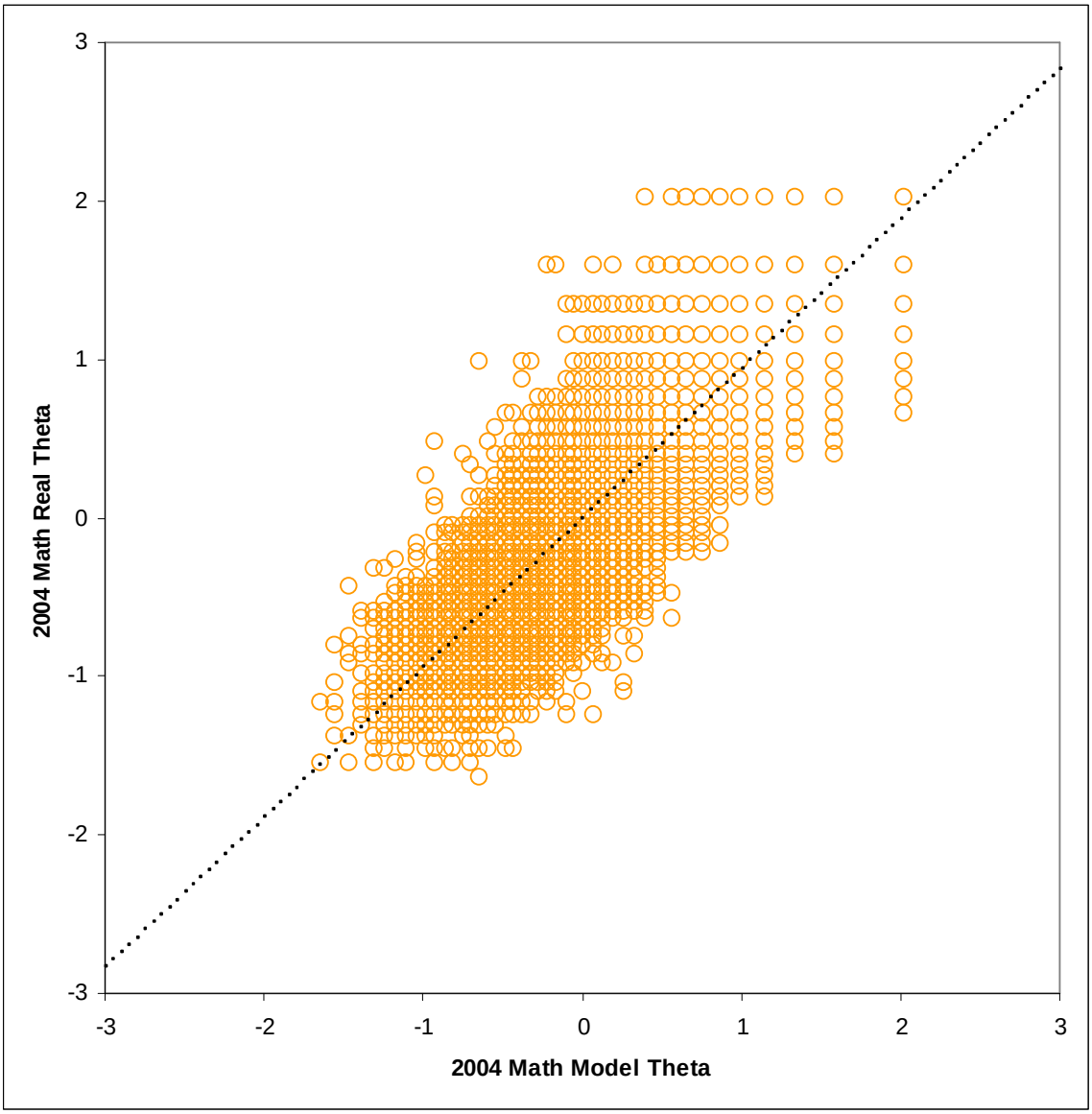

Figure 6.5 Graph of 2004 Math real world θ as a function of the model θ value (Fit Line: y = 1.119, R^2 = 0.712)

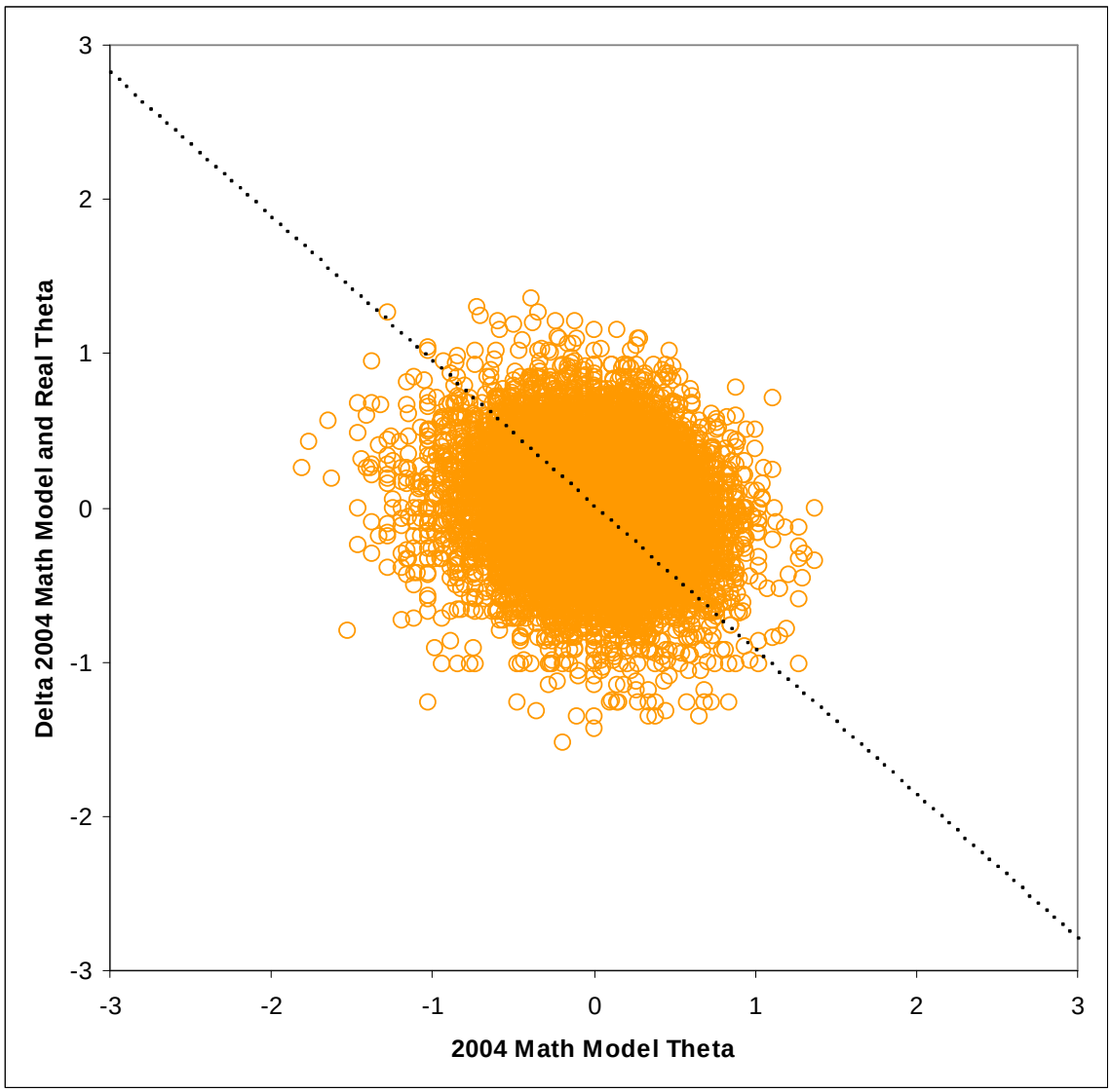

Figure 6.6 Difference between 2005 Math model and real θ value as a function the difference between 2004 Math model and real θ
(Fit Line: y = -0.936x + 0.009, R^2 = 0.022)

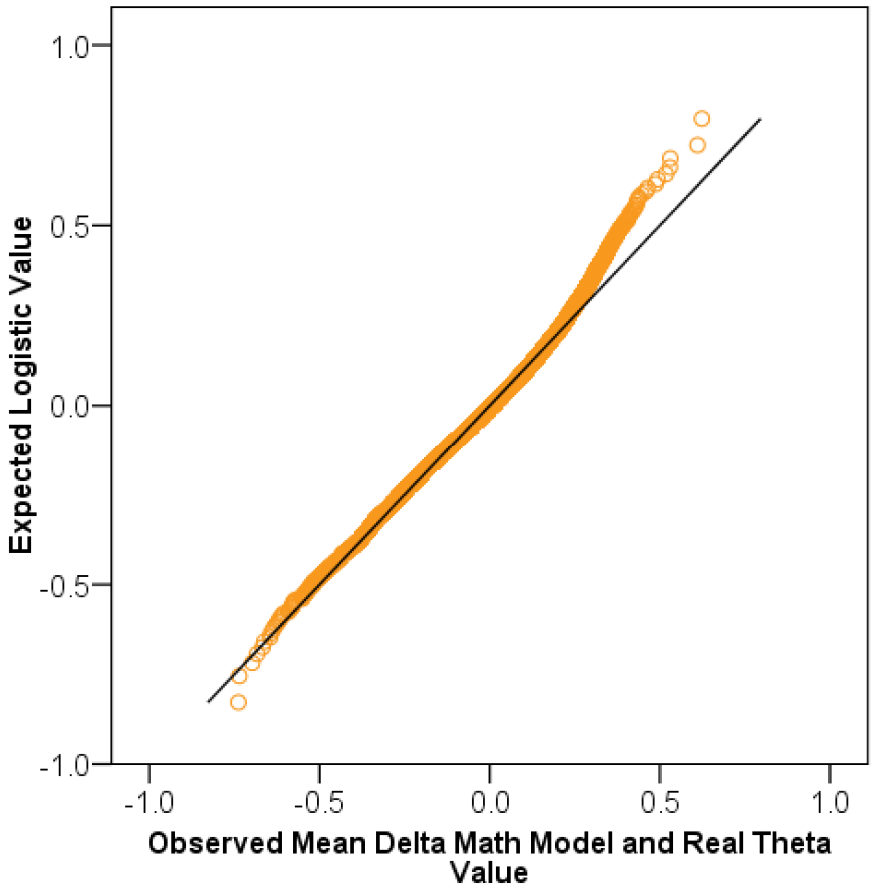

Figure 6.7 Q-Q plot of the mean Math difference between model and real world θ values across years

Discussion

Chapter 4 dealt with the trends observed in the real world data. **Chapter 5** showed how the agent based computer model of IRT-1PL was able to mimic these trends in the agents. In this chapter, we have attempted to verify that the trends in the real world can be directly attributed to the IRT-1PL theoretical testing framework by using the estimated domain latent trait values from real world students in the model. The results show that while there is a slight deviation from a perfect IRT-1PL execution in the real world, the model is able to produce data output that is uncannily close to that of the real world. This serves to first validate the model, and secondly to confirm that the results of the TAKS exams are almost exclusively due to IRT-1PL.

The fact that the model predicts θ values that are just as well and often times better correlated to the various sections of the TAKS exam than the TAKS scale scores themselves make them practically indistinguishable from the scale scores without doing statistical analysis on the student dataset. For instance in **Table 6.4**, the model M04 θ values are 84% correlated to the real M06 θ values. The real M04 θ values are only 76% correlated to the real M06 θ values. Also the model θ values can explain over 70% of the variance in the real θ values. Based on **Chapter 5**, the remaining less than 30% of the variance is mostly due to random measurement error. Compare this to the report TEA released in 2005 stating that the Algebra I grade given by teachers to students could only explain 35% of the variance in the End-of-Course (EOC) Algebra I exam scores (TEA

Division of Student Assessment, 2005). While the EOC exams are not the same as the TAKS exam, they are built using IRT-1PL just like the TAKS exam. They should therefore exhibit similar behaviors. The fact that the model can explain more of the variance in test scores than the grades given to students by teachers is disconcerting.

CHAPTER 7: Conclusions and Consequences

The goal of education is to empower students with the skills and knowledge necessary to succeed in life. An equitable educational system allows students of any background to succeed by allowing them to learn and grow to their own full potential. The goal of testing is to evaluate whether students have met the standards in the various domains that experts in those domains consider to be acceptable. It is important then to ensure that our tests do what they are supposed to do. In this dissertation, we have used two methods to test whether the TAKS exam measures student achievement in four different domains. First we used real world data obtained from TEA. Then we used a computer model generated based on the theoretical framework that was used to make the TAKS exam. Then we combine the two together to both validate the model and show that the trends seen in the TAKS scores are mostly likely a direct result of the IRT-1PL framework. When all the results are taken together, they provide compelling evidence that the TAKS exam has failed its intended purpose of evaluating student achievement at an abysmal level and worse, it maintains the existing inequities in students.

Conclusions

Having analyzed data from the real world and the agent based computer model, certain conclusions can now be drawn definitively about the TAKS exam. They are as follow:

- Any changes in student scores across years are purely a result of regression towards the mean. No meaningful change was observed in the real world data.
- The consistency of student scores across domains indicates that the domains tested for very little content material.
- Students are being rank ordered persistently across years and consistently across domains based on an uncharacterized mental latent trait.
- Based on these conclusions, the TAKS exam must be instructionally insensitive.

The fact that any changes in student scores across years are due to regression towards the mean indicates that the students are maintaining their relative position to each other on the domain θ scales, i.e. they are not mobile. Any observed changes are due to random measurement error and so the students are merely stochastically fluctuating about their true θ value. This would not be incriminating to the TAKS exam if the θ scales for each domain were orthogonal or at least did not share a disproportionately large amount of variance. If we

assume that each domain possesses unique knowledge and skills that are not present in the other domains, then the fact that each of the domain θ scales share so much of the variance indicates the existence of a common latent trait unrelated to the content in the individual domains. We have labeled this common latent trait the profiling mental latent trait (LTP). The persistence of the rank ordering of students across years and domains indicates that LTP is an extremely stable mental latent trait in students and therefore insensitive to instruction.

When taken as a whole, the bottom line is that the TAKS exam rank orders students on a persistent latent trait that we have called LTP, and that it is insensitive to instruction and therefore does not measure student achievement as is assumed by the stakeholders, as mandated by the law, and as claimed by the test developers. No doubt these conclusions will draw a lot of interest from both the proponents and opponents of standardized testing due to the implications.

Causes of Failure

To answer this question of why the TAKS exam has failed requires us to thoroughly examine the assumptions of IRT. A test is only as good as the foundation from which it is built and the TAKS exam is built on IRT-1PL. One of the primary assumptions of IRT is that it measures a unidimensional latent trait. IRT does not explicitly define the nature of this latent trait but rather uses the set of items during test calibration to derive a common latent trait across all items relative to the test calibration sample. What the items have in common is assumed to be what we want to test: the specific content objectives and is based on the face validity of the items. However, also present in the items are other factors that affect how students respond to it aside from the specific content objectives. These other factors collectively constitute LTP and it is probable that LTP stays the consistent across domains since it is not content specific. Now if we assume that the specific content objective also stays the same across items, then the b-value is a difficulty measure of the combination of specific content objectives and LTP. However, the set of items used during test calibration are actually not designed to test for just one specific content objective, but rather a set of content objectives for the whole domain. **Table 7.1** shows the number of main objectives and sub-objectives on each section of the 2006 TAKS exam (TEA Division of Student Assessment, 2006). Since test calibration uses the entire set of items to measure what the items have in common relative to the test calibration sample, the b-value ironically would not represent any of the specific

content objectives since they are not shared across all items except at the most basic level of the domain. On the other hand, the collective factors that we have labeled as LTP probably remain common in all items. This is most likely the reason why the TAKS exam seems to be instructionally insensitive, and rank orders students on LTP across domains.

In order for IRT to measure student achievement, we would have to create individual exams for each of the objectives. In this manner, we can ensure that the items now share that one objective across all items. However, this is unpractical in the real world. The math section contains 65 objectives. Assuming a minimum of ten items per objective to reasonably gauge achievement, that would be a total of 650 items for the math section alone! Furthermore, there would be a total of 65 scores in the math domain: one for each objective. While this might actually be useful to teachers since it would indicate which objectives students understand and which they need more work on, the trend is to gravitate towards one score to cover all of the objectives so as to make for easier interpretation. It is statistically unsound to average all of the scores together since each score is a different nominal category. The desire for one generic score on each domain is the ultimate undoing of IRT as a theoretical testing framework to be used to measure student achievement.

Another major assumption of IRT is that the item parameters are invariant. Items cannot change their b-value and items are selected specifically to remain stable across populations. This creates a bias against items that tend to have their b-values change. One of the threats to item invariance is instructional

sensitivity. Items for which students can learn means that their relative ranking can change, but such changes invalidates the external θ scale that is supposed to be objective. This means that items are selected to be instructionally insensitive so as to allow for the conservation of the θ scale. This is contrary to what a measurement of achievement would look like: one that *is* instructionally sensitive. While the invariance of the external θ scale allows for comparisons to be made across years, it is a reason why the TAKS exam cannot measure student achievement since achievement should be variable from year to year with different groups of students.

The very basic assumptions of IRT are untenable to an authentic assessment of student achievement. The desire for one generic score in each domain undermines the specific content objectives that students are expected to achievement while the desire to be able to compare students across years undermines changes in student achievement from year to year. Perhaps it would be wise then to stop standardized testing before more damage is accumulated while newer and more authentic forms of assessments are developed.

Domain	Total Number of Main Objectives	Total Number of Sub Objectives	Number of Items
Reading	6	42	48
Mathematics	10	65	60
History	5	63	55
Science	5	36	55

Table 7.1 Number of objectives in each domain of the 2006 TAKS exam

Consequences

Considering the importance of high stakes standardized testing in the current educational environment in the United States, it is imperative that we, as stakeholders in the educational system, understand how standardized testing works. In this dissertation, we have focused solely on the TAKS exam as a prototypical standardized IRT test. However, the results should extend to other standardized tests built using IRT. As such, the implications of this dissertation should extend to a national, and indeed, international level.

Across the state of Texas, teachers and administrators are working to educate our students. NCLB was designed to hold them accountable for the success of our students. There is no doubt that teachers need to be held accountable for their practice so as to ensure the highest standards of teaching for our students. However, the measures for accountability should not be based on the TAKS exam. Teachers and schools that do not meet the requirements of AYP are sanctioned for their supposed failures. This sanctioning is unfair since the results of the TAKS exam are independent of student achievement. Rather, teachers are being sanctioned for having a certain profile of students in their classroom. While the profile is not well characterized, it is known to include factors such as ethnicity, socioeconomic status, and gender. Students with the profiles that generate low TAKS scores are usually the one most in need of intervention (ethnic minority, low SES, etc). Sanctioning teachers for students of

this profile will only relegate these students to the worst standards of teaching possible.

Also, many different programs of educational research use standardized test scores as a benchmark of success (Stroup et al, 2007). We have shown how this would be a pitfall since increases in achievement cannot be measured using these exams. If we are to have meaning interventions for our students, we need to move away from using standardized test scores as our benchmarks. Standardized tests simply do not test for real learning and one has to question how many great interventions were discarded due to lack of promise based only on standardized test scores. Furthermore, even if a meaningful intervention manages to be able to improve standardized test scores, the allowable improvements are incredibly small. Our analysis in this dissertation indicates that approximately 10% more of the variance is available before the theoretical limits of IRT-1PL are reached. This is using a generous model that does not account for student guessing. If the model did account for guessing, the allowable improvements are most likely to be less than 10%.

The lack of congruence between what is being taught and what is being assessed is disturbing. Considering the lack of utility that standardized testing possesses, it is perhaps time to stop them. The damage they can cause is greater than any perceived value they may have. These damages include affective issues with students who perform poorly, the maintenance of the “achievement” gap, and the sanctioning teachers for having the wrong profile of students in their classrooms. The stated intention of NCLB is to close the

achievement gap and raise the bar on academic achievement in U.S. students through accountability on the part of students, teachers, parents, and school systems. This goal is completely contrary to what high stakes standardized testing is actually accomplishing. As W. James Popham says, accountability only works if high stakes standardized tests actually measures the outcome of instruction (achievement) and therefore is a measure of instructional quality (2007).

Further Study

One of the things not discussed yet is guessing on the TAKS since this dissertation is only concern with modeling the theoretical foundation of the TAKS exam, IRT-1PL. The TAKS exam employs mostly multiple choice items with four choices to determine student scores. This means that if a student simply guesses randomly on the exam, they will have a 25% chance of responding to an item correctly. This would change the lower asymptote of the IRFs. A perusal of the agent base model will indicate that a guessing parameter was actually included in the code but not used in this dissertation, and that preliminary work with it has indicated that when students guess at a 25% rate, the differences between the real world and the model practically vanish. Since this is only a preliminary exploration, more work needs to be done before a conclusion can be drawn definitively.

Glossary

1PL – One Parameter Logistic model
3PL – Three Parameter Logistic Model
Adequate Yearly Progress – The requirement that teachers and schools make improvements in their scores each year or else face sanctioning.
AYP – See Adequate Yearly Progress
b-value – The difficulty parameter of an item
Classical Testing Theory –Formerly predominant testing framework that was dependent on populations that are normally distributed
CTT – See Classical Testing Theory
Domain – One of the four subject areas tested for on the TAKS exam: Reading, Math, History, and Science.
Domain latent trait – The statistical construct that represents the students' achievement level within that domain and thus is what the TAKS exam is measuring.
IRF – See Item Response Function.
IRT – See Item Response Theory
Item Response Function – The probability function of an item at a specific b-value relative to the entire population.
Item Response Theory – Theoretical testing framework that relies on intrinsic parameters for students and items, obviating the need for specific population distributions.
Maximum Likelihood Estimation – Iterative estimation process used to determine student θ and item b-values in this dissertation.
MLE – See Maximum Likelihood Estimation.
NCLB – See No Child Left Behind Act.
No Child Left Behind Act – Federal law passed in 2001 mandating that every state administer a standardized exam to hold students, teachers, and school systems accountable as well as mandating that schools must show "Adequate Yearly Progress".
PARAM-1PL – Freeware computer program to estimate IRT parameters made available by Lawrence Rudner.
Pearson Educational Mearurement – Private contractor to TEA responsible for the development and administration of the TAKS exam.
PEM – See Pearson Educational Measurement.
Rasch Partial Credit Model – IRT based model to grade multiple response type items
RPCM – See Rasch Partial Credit Model
SEM – See Structural Equation Modeling
Structural Equation Modeling – Statistical analysis that determines causality based on the fitting of data to a structural model.
TAKS – See Texas Assessment of Knowledge and Skills.
TEA – See Texas Education Agency.

Texas Assessment of Knowledge and Skills – State of Texas's standardized summative exam used to comply with NCLB.

Texas Education Agency – State of Texas's governmental branch tasked with regulating the state's educational system.

Theta or **θ value** – Ability parameter of examinees

Appendix A: Codes for the NetLogo TAKS IRT-1PL model

Text in grey indicate they are comments by the author to explain the codes

```
globals ; globals are the global variables used in the model
        [ ; opening bracket
        posit ; posit is a simple counter used to graph the cumulative distribution
        turtle-population ; number of turtles in model, the bigger the number the better the
        ; estimation process will be for theta and item parameter estimation
        time ; a counter to keep track of time in terms of numbers of cycles the routine runs
        Mean-LT ; mean value of a latent trait, used to standardize scores
        SD-LT ; Standard deviation of a latent trait, used to standardize scores

        ; the following are the list of b-values for the specific section and year - these values were
        ; derived from the Analytic dataset from actual students taking the TAKS exam
        R04-b-var-list ; Reading 2004
        R05-b-var-list ; Reading 2005
        R06-b-var-list ; Reading 2006
        M04-b-var-list ; Math 2004
        M05-b-var-list ; Math 2005
        M06-b-var-list ; Math 2006
        H05-b-var-list ; History 2005
        H06-b-var-list ; History 2006
        S05-b-var-list ; Science 2005
        S06-b-var-list ; Science 2006
        ] ; closing bracket

turtles-own ; these are the variables that each individual turtle has, and whose values are
; independent of other turtles
        [ ; opening bracket

        ; the following are the actual latent traits
        LTP ; profiling latent trait
        LTR ; reading latent trait
        LTM ; math latent trait
        LTH ; history latent trait
        LTS ; science latent trait

        ; the following are the binary variables to indicate whether a turtle has been "infected"
        ; with the specific latent trait yet. 0 = uninfected and 1 = infected
        lightbulb-LTP
        lightbulb-LTR
        lightbulb-LTM
        lightbulb-LTH
        lightbulb-LTS

        Ans-list ; variable to temporary store the answer list to each section as the turtles take the
        ; exam before being offloaded on the actual variable list below

        ; the following are the list of answers to each section and year for the turtles (the
        ; response set of each turtle)
        R04-Ans ; turtle's response set to 2004 Reading
        R05-Ans ; turtle's response set to 2005 Reading
```

```
        R06-Ans ; turtle's response set to 2006 Reading
        M04-Ans ; turtle's response set to 2004 Math
        M05-Ans ; turtle's response set to 2005 Math
        M06-Ans ; turtle's response set to 2006 Math
        H05-Ans ; turtle's response set to 2005 History
        H06-Ans ; turtle's response set to 2006 History
        S05-Ans ; turtle's response set to 2005 Science
        S06-Ans ; turtle's response set to 2006 Science

        ; the following are the turtle's raw score in each section and year
        R04-Raw ; turtle's raw score to 2004 Reading
        R05-Raw ; turtle's raw score to 2005 Reading
        R06-Raw ; turtle's raw score to 2006 Reading
        M04-Raw ; turtle's raw score to 2004 Math
        M05-Raw ; turtle's raw score to 2005 Math
        M06-Raw ; turtle's raw score to 2006 Math
        H05-Raw ; turtle's raw score to 2005 History
        H06-Raw ; turtle's raw score to 2006 History
        S05-Raw ; turtle's raw score to 2005 Science
        S06-Raw ; turtle's raw score to 2006 Science

        ; the following are the domain raw scores of the turtle
        Reading-Raw ; turtle's total Reading raw score
        Math-Raw ; turtle's total Math raw score
        History-Raw ; turtle's total History raw score
        Science-Raw ; turtle's total Science raw score
        ] ;closing bracket

to setup ; this is the setup routine to start the model by setting up the turtle population's initial
; values
        clear-all ; resets the model and clears all variables and turtles from any previous runs
        set turtle-population 30000 ; allows us to flexibly rescale model population size as to
        ; whatever number desire, default is 30,000
        create-turtles turtle-population ; creates the number of turtles designated in turtle-
        ; population

        ask turtles
                [
                set hidden? True ; hides the turtles, allows the model to run faster since there is
        ; no visual graphics processing
                ]

        spread ; runs the spread routine

end ; end of the setup routine

to spread ; routine to distribute turtles randomly on map and catch the latent trait "disease"

; This routine allows each latent trait to be distributed as if they were a contagion spread by
; physical contact. This biologically derived way of spreading the latent trait yields a more
; "natural" population based on the logistic growth curve. Using it in this IRT makes sense since
; we are trying to create an ecologically valid model. The reason we have to run the spreading
; routine for each individual latent trait separately is because having them run together created
; correlations between them due to localized population interaction and distribution effects. By
```

```
; randomly spreading them out after each run, the latent traits are ensured to be completely
; orthogonal. This is in keeping with IRT's unidimensional aspect of the latent traits

        ; The following code is for the Profiling Latent Trait. The codes for the other latent traits
        ; are the same for the most part aside from the names. Any new code will be commented
        ; as they occur
        ask turtles
                [
                setxy round random-pxcor round random-pycor ; randomly distributes turtles on
                ; to a patch
                set heading ((random 4) * 90) ; sets each turtle's heading randomly in one of the
                ; four cardinal directions
                ]

        ask one-of turtles
                [
                set lightbulb-LTP 1 set LTP time ; this is to initially "infect" one of the turtles so
                ; that the LT can be passed on
                ]

        while [(count turtles with [lightbulb-LTP = 1] < turtle-population)] ; keeps the LT spreading
        ; until all turtles are "infected"
                [
                set time time + 1 ; advances the time counter by one each cycle
                ask turtles
                        [
                        forward 1 ; turtles take a step forward
                        set heading ((random 3 - 1) * 90) ; and turns randomly on a straight
                        ; ahead, right, or left direction. Turtles do not turn backwards.
                        if lightbulb-LTP = 0 and any? other turtles-here with [lightbulb-LTP = 1]
                        ; if turtle is "uninfected" asks if there are any other
                                [
                                set lightbulb-LTP 1 set LTP time ; turtle on patch who is
;                               "infected" so that the turtle can get "infected"
                                ]
                        ]
                ]

        ; This is for the Reading Latent Trait
        ask turtles
                [
                setxy round random-pxcor round random-pycor
                set heading ((random 4) * 90)
                set time 0 ; resets timer from last spreading of LT
                ]

        ask one-of turtles
                [
                set lightbulb-LTR 1 set LTR time
                ]

        while [count turtles with [lightbulb-LTR = 1] < turtle-population]
                [
                set time time + 1
```

```
        ask turtles
                [
                forward 1
                set heading ((random 3 - 1) * 90)
                if lightbulb-LTR = 0 and any? other turtles-here with [lightbulb-LTR = 1]
                        [
                        set lightbulb-LTR 1 set LTR time
                        ]
                ]
        ]

; This is for the Math Latent Trait
ask turtles
        [
        setxy round random-pxcor round random-pycor
        set heading ((random 4) * 90)
        set time 0
        ]

ask one-of turtles
        [
        set lightbulb-LTM 1 set LTM time
        ]

while [count turtles with [lightbulb-LTM = 1] < turtle-population]
        [
        set time time + 1
        ask turtles
                [
                forward 1
                set heading ((random 3 - 1) * 90)
                if lightbulb-LTM = 0 and any? other turtles-here with [lightbulb-LTM = 1]
                        [
                        set lightbulb-LTM 1 set LTM time
                        ]
                ]
        ]

; This is for the History Latent Trait
ask turtles
        [
        setxy round random-pxcor round random-pycor
        set heading ((random 4) * 90)
        set time 0
        ]

ask one-of turtles
        [
        set lightbulb-LTH 1 set LTH time
        ]

while [count turtles with [lightbulb-LTH = 1] < turtle-population]
        [
        set time time + 1
```

```
		ask turtles
			[
			forward 1
			set heading ((random 3 - 1) * 90)
			if lightbulb-LTH = 0 and any? other turtles-here with [lightbulb-LTH = 1]
				[
				set lightbulb-LTH 1 set LTH time
				]
			]
		]

	; This is for the Science Latent Trait
	ask turtles
		[
		setxy round random-pxcor round random-pycor
		set heading ((random 4) * 90)
		set time 0
		]

		ask one-of turtles
			[
			set lightbulb-LTS 1 set LTS time
			]
		while [count turtles with [lightbulb-LTS = 1] < turtle-population]
			[
			set time time + 1
			ask turtles
				[
				forward 1
				set heading ((random 3 - 1) * 90)
				if lightbulb-LTS = 0 and any? other turtles-here with [lightbulb-
LTS = 1]
					[
					set lightbulb-LTS 1 set LTS time
					]
				]
			]

	set-current-plot "Population Distribution" ; selects the plot to start plotting
	while [posit < time]  ; keeps plotting until all students are plotted
		[
		set posit posit + 1 ; advances posit by one each cycle
		set-current-plot-pen "Cumulative Distribution" ; sets the plot pen
		plot (count turtles with [LTP < posit] / turtle-population * 100) ; plots the
		; cumulative distribution
		]

	; the following code standardizes the LTP to a z-score. Codes are the same for the other
	; LT's aside from name changes.
	set Mean-LT mean [LTP] of turtles ; sets the Mean-LT variable to the mean LTP value
	set SD-LT standard-deviation [LTP] of turtles ; sets the SD-LT to the standard deviation
	; of the LTP value
	ask turtles
		[
```

```
        set LTP (LTP - Mean-LT) / SD-LT ; z-score transformation
        ]

set Mean-LT mean [LTR] of turtles
set SD-LT standard-deviation [LTR] of turtles
ask turtles
        [
        set LTR (LTR - Mean-LT) / SD-LT
        ]

set Mean-LT mean [LTM] of turtles
set SD-LT standard-deviation [LTM] of turtles
ask turtles
        [
        set LTM (LTM - Mean-LT) / SD-LT
        ]

set Mean-LT mean [LTH] of turtles
set SD-LT standard-deviation [LTH] of turtles
ask turtles
        [
        set LTH (LTH - Mean-LT) / SD-LT
        ]

set Mean-LT mean [LTS] of turtles
set SD-LT standard-deviation [LTS] of turtles
ask turtles
        [
        set LTS (LTS - Mean-LT) / SD-LT
        ]

; These are the codes used to set the linkage between the domain latent traits and the
; profiling latent trait from the slider on the user interface. It uses a partial shared variance
; system as one would use in determining shared variances in SEM
ifelse Sim-Link = false ; this allows the linkage codes to be run as set by user or as a
; simulation of the actual data from the real world SEM.
        [
        ask turtles
                [
                set LTR ((1 - LTR-Link) * LTR) + (LTR-Link * LTP)
                set LTM ((1 - LTM-Link) * LTM) + (LTM-Link * LTP)
                set LTH ((1 - LTH-Link) * LTH) + (LTH-Link * LTP)
                set LTS ((1 - LTS-Link) * LTS) + (LTS-Link * LTP)
                ]
        ]

        ; Same as above except use these linkage codes to simulate the actual data
        ; instead of the above codes so that each specific latent trait is linked to the
        ; profiling trait to specifically simulate real world data.
        [
        ask turtles
                [
                set LTR ((1 - 0.622) * LTR) + (0.622 * LTP)
                set LTM ((1 - 0.665) * LTM) + (0.665 * LTP)
```

```
                set LTH ((1 - 0.744) * LTH) + (0.744 * LTP)
                set LTS ((1 - 1.000) * LTS) + (1.000 * LTP)
                ]
        ]

; the following code standardizes the LTP to a z-score. Codes are the same for the other
; LT's aside from name changes.
set Mean-LT mean [LTP] of turtles ; sets the Mean-LT variable to the mean LTP value
set SD-LT standard-deviation [LTP] of turtles ; sets the SD-LT to the standard deviation
; of the LTP value
ask turtles
        [
        set LTP (LTP - Mean-LT) / SD-LT ; z-score transformation
        ]

set Mean-LT mean [LTR] of turtles
set SD-LT standard-deviation [LTR] of turtles
ask turtles
        [
        set LTR (LTR - Mean-LT) / SD-LT
        ]

set Mean-LT mean [LTM] of turtles
set SD-LT standard-deviation [LTM] of turtles
ask turtles
        [
        set LTM (LTM - Mean-LT) / SD-LT
        ]

set Mean-LT mean [LTH] of turtles
set SD-LT standard-deviation [LTH] of turtles
ask turtles
        [
        set LTH (LTH - Mean-LT) / SD-LT
        ]

set Mean-LT mean [LTS] of turtles
set SD-LT standard-deviation [LTS] of turtles
ask turtles
        [
        set LTS (LTS - Mean-LT) / SD-LT
        ]

; rescales the z transformed latent traits to the LT-Mean and LT-SD sliders on the user
; interface or to real world population values
ifelse Sim-Pop = false
        [
        ask turtles
                [
                set LTR LTR * LT-SD + LT-Mean
                set LTM LTM * LT-SD + LT-Mean
                set LTH LTH * LT-SD + LT-Mean
                set LTS LTS * LT-SD + LT-Mean
                ]
```

```
                ]

                ; same as above but these are the values needed to simulate the real world
                [
                ask turtles
                        [
                        set LTR LTR * 0.780 + 0.777
                        set LTM LTM * 0.898 + 0.395
                        set LTH LTH * 0.915 + 0.681
                        set LTS LTS * 0.780 + 0.325
                        ]
                ]

        set-current-plot "Histogram" ; selects the plot for plotting
        set-histogram-num-bars 20 ; sets the number of bars on the histogram
        histogram [LTP] of turtles ; generate a histogram of the turtles' LTP

end ; end of spread routine

to go ; the go routine, activated on the front user interface

        ; the following codes populate the TAKS section list with their respective internal b-values
        ; as determine from the Analytic dataset
        set R04-b-var-list
                [
                -0.544 -0.985 -1.163 -0.668 -0.807 -0.651 -0.67 -0.577 -0.417 -1.225 -
1.778 -0.625 -0.861 -1.09 -0.676 0.066 -1.403 -1.799 -1.244 -0.809 -
0.389 -1.712 -1.321 -0.923 -0.913 -0.768 -1.243 -0.86 -1.009 0.115 -1.468 -0.226 0.368
                ]

        set R05-b-var-list
                [
                -1.427 -1.538 -1.084 -0.642 -1.269 -2.058 -0.813 -1.354 -1.914 -2.155 -
1.049 -0.844 -1.192 -1.264 -1.162 -0.436 -0.714 -1.037 -0.688 -1.211 -
1.14 -1.405 -1.089 0.219 -0.782 -0.856 -0.622 -0.596 -2.434 -0.952 -2.215 -1.527 -1.467 -
0.805 -2.396 -1.332 -1.326 -0.978 -2.247 -0.824 -0.615 -1.913 -1.911 -
1.493 -0.834 -1.49 -1.258 -1.39
                ]

        set R06-b-var-list
                [
                -1.16 -1.154 -1.508 -2.442 -1.142 -1.413 -1.603 -2.023 -1.509 -1.717 -
2.274 -1.615 -1.494 -1.2 -1.356 -1.61 -1.337 -1.642 -1.357 -1.248 -1.495 -
1.227 -1.454 -0.796 -1.168 -1.52 -1.038 -1.83 -1.072 -2.141 -2.28 -0.708 -1.81 -2.119
                -0.512 -1.67 -1.556 -1.195 -1.468 -1.499 -1.842 -1.215 -0.919 -0.868 -1.91
-1.759 -0.733 -1.364
                ]

        set M04-b-var-list
                [
                -1.714 -1.612 -0.933 -1.104 -1.055 -1.044 -1.023 0.161 -0.682 -0.8 -0.467
-0.935 -0.851 -0.467 -0.181 -0.266 -0.107 -0.068 -0.145 -0.264 -0.13 -
0.097 -0.197 -0.031 -0.698 0.218 0.548 -0.007 0.172 0.042 0.022 -0.081 -0.232 -0.008
```

-0.143 -0.707 -0.242 -0.304 -0.372 -0.97 -0.538 -0.572 -0.594 -0.378 -0.827 -1.062 -0.78 0.621 -0.743 -1.142 -1.139 -1.543
]

set M05-b-var-list
[
-1.151 -0.755 -1.354 -1.566 -0.547 -0.661 -1.14 -0.696 -0.712 -0.882 -0.402 -0.69 -0.081 -1.575 -0.957 -0.038 -0.56 -0.729 0.008 -0.526 -0.904 0.159 0.066 0.033 -1.066 -1.003 0.284 0.099 -0.134 -0.012 0.322 -0.071 0.377 -0.481 -0.727 -0.64 -0.731 0.317 -1.112 0.117 0.089 -1.014 -0.132 -0.198 0.383 -0.127 0.059 -0.449 -1.11 -0.843 -0.142 -0.46 -0.929 -0.813 -0.71 -1.673
]

set M06-b-var-list
[
-1.608 -1.091 -1.431 -0.945 -0.836 -0.684 -1.129 -0.977 -1.163 -0.604 -1.268 -0.673 -0.275 -1.571 -1.023 -0.466 -0.337 -1.045 0.337 -0.056 -0.528 -0.163 -0.198 -0.737 -0.267 -1.432 -0.245 -0.537 0.306 -0.565 -0.086 -0.156 0.063 0.225 0.258 -0.015 -0.462 -0.143 -0.743 -1.405 -0.646 -0.793 -0.617 0.214 -0.602 -1.374 0.052 -1.454 -0.36 -0.616 -0.13 0.168 -1.169 -1.029 -1.016 -0.587 -0.466 -1.122 -0.889 -1.055
]

set H05-b-var-list
[
-2.02 -1.737 -1.469 -1.648 -0.43 -1.288 -1.285 -1.495 -1.255 -1.249 -1.039 -0.946 -0.481 -1.525 -1.171 -0.864 -0.885 -0.598 -0.8 -0.664 -0.108 -0.182 -1.459 -0.951 -0.392 -0.262 -0.441 0.06 -0.11 -0.302 -0.453 -0.574 -0.32 -0.379 -0.846 -0.946 -0.366 -1.002 -0.982 -1.011 -0.579 -1.276 -0.998 -1.082 -0.573 -1.234 -0.877 -0.717 -1.316 -1.317
]

set H06-b-var-list
[
-2.236 -2.206 -1.834 -1.639 -1.996 -1.53 -1.583 -0.984 -1.143 -1.153 -0.38 -1.093 -0.827 -0.807 -1.122 -0.213 -0.764 -0.828 -0.212 -0.428 -0.149 -0.427 -0.437 -0.372 -0.817 0.014 -0.134 -1.166 -0.711 -0.145 -0.545 -0.537 -0.553 -0.219 -1.037 -0.866 -0.465 -0.243 -0.173 -0.777 -0.863 -0.945 -0.828 -1.064 -0.907 -0.991 -1.052 -1.455 -1.276 -1.557 -0.694 -1.202 -2.046 -1.966 -1.563
]

set S05-b-var-list
[
-1.608 -1.814 -1.5 -1.182 -0.773 -0.796 -0.666 -0.552 -0.442 -0.575 -0.784 -0.511 -0.541 -0.144 0.027 -0.72 -0.346 -0.276 0.293 -0.922 0.437 0.319 0.123 0.345 0.099 0.167 0.06 -0.054 -0.021 -0.084 -0.089 -0.069 -0.321 -0.142 -0.292 -0.187 -0.277 -0.337 -0.43 0.016 -0.482 -0.518 -0.555 -0.467 -0.435 -0.84 -0.613 -0.438 -0.723 -0.773 -0.963 -1.01 -1.228 -1.142 -1.634
]

set S06-b-var-list

```
                [
                -1.13  -1.428  -1.173  -1.356  -1.098  -0.869  -0.186  -0.175  0.202  -0.786  0.636
-0.002  -0.972  -0.769  -0.039  -0.423  0.106                    -0.511  -0.933  -0.88  0.278  -
0.066  -0.43  -0.86  -0.289  -0.246  -0.478  -0.439  -0.2  -0.436  0.195  -0.516  -0.671  -0.118  -
0.226                -0.888  0.058  0.005  -0.729  -0.929  0.03  -0.177  0.152  -0.606  -0.253  -
0.292  -0.788  0.392  0.12  -0.976  -0.443  -1.305  -0.504                -1.436  -1.352
                ]

        test ; runs the test routine

end ; end of the go routine

to test ; the test routine
; This is the bulk of the IRT modeling code. Until now, it was merely to setup the inital population
; to be tested. The way it works is given the a priori determined latent trait value from above, what
; is the probability of the turtle getting an item right based on each item's b-value. Then a random
; number generator is used to create a random value from 0 to 1. If the number is lower than the
; probability given by the IRT equation the turtle got the item correct and is listed as "True". If the
; number is larger, then the turtle got the item wrong and is listed as "False". This is in keeping
; with the idea in IRT that b-values are related to student's probability of getting an item right and
; so using the random number generator makes sense. Notice that a guessrate parameter was
; added in the code. Even though the point is to model the IRT-1PL framework of the TAKS exam,
; it is silly to believe that students do not guess on the exam, especially when there are only four
; choices to chose from.

        foreach sort turtles ; ask each turtle to individually take the exam in order
                [
                ask ?
                        [
                        set R04-Ans (map [(Guess-Rate + ((1 - Guess-Rate) / ( 1 + e ^ (-1 * (LTR
- ?)))))) > (random 100 * 0.01)] R04-b-var-list)
                        set R05-Ans (map [(Guess-Rate + ((1 - Guess-Rate) / ( 1 + e ^ (-1 * (LTR
- ?)))))) > (random 100 * 0.01)] R05-b-var-list)
                        set R06-Ans (map [(Guess-Rate + ((1 - Guess-Rate) / ( 1 + e ^ (-1 * (LTR
- ?)))))) > (random 100 * 0.01)] R06-b-var-list)
                        set M04-Ans (map [(Guess-Rate + ((1 - Guess-Rate) / ( 1 + e ^ (-1 *
(LTM - ?)))))) > (random 100 * 0.01)] M04-b-var-list)
                        set M05-Ans (map [(Guess-Rate + ((1 - Guess-Rate) / ( 1 + e ^ (-1 *
(LTM - ?)))))) > (random 100 * 0.01)] M05-b-var-list)
                        set M06-Ans (map [(Guess-Rate + ((1 - Guess-Rate) / ( 1 + e ^ (-1 *
(LTM - ?)))))) > (random 100 * 0.01)] M06-b-var-list)
                        set H05-Ans (map [(Guess-Rate + ((1 - Guess-Rate) / ( 1 + e ^ (-1 * (LTH
- ?)))))) > (random 100 * 0.01)] H05-b-var-list)
                        set H06-Ans (map [(Guess-Rate + ((1 - Guess-Rate) / ( 1 + e ^ (-1 * (LTH
- ?)))))) > (random 100 * 0.01)] H06-b-var-list)
                        set S05-Ans (map [(Guess-Rate + ((1 - Guess-Rate) / ( 1 + e ^ (-1 * (LTS
- ?)))))) > (random 100 * 0.01)] S05-b-var-list)
                        set S06-Ans (map [(Guess-Rate + ((1 - Guess-Rate) / ( 1 + e ^ (-1 * (LTS
- ?)))))) > (random 100 * 0.01)] S06-b-var-list)
                        ]
                ]

        ; The following codes changes the response set from "True" and "False" to 1 and 0
        ; respectively so that PARAM-1PL can read the response set.
```

```
ask turtles
        [
        set Ans-list (map [ifelse-value (? = true) [1][0]] R04-Ans)
        set R04-Ans Ans-list
        set R04-Raw length (remove 0 Ans-list)

        set Ans-list (map [ifelse-value (? = true) [1][0]] R05-Ans)
        set R05-Ans Ans-list
        set R05-Raw length (remove 0 Ans-list)

        set Ans-list (map [ifelse-value (? = true) [1][0]] R06-Ans)
        set R06-Ans Ans-list
        set R06-Raw length (remove 0 Ans-list)

        set Ans-list (map [ifelse-value (? = true) [1][0]] M04-Ans)
        set M04-Ans Ans-list
        set M04-Raw length (remove 0 Ans-list)

        set Ans-list (map [ifelse-value (? = true) [1][0]] M05-Ans)
        set M05-Ans Ans-list
        set M05-Raw length (remove 0 Ans-list)

        set Ans-list (map [ifelse-value (? = true) [1][0]] M06-Ans)
        set M06-Ans Ans-list
        set M06-Raw length (remove 0 Ans-list)

        set Ans-list (map [ifelse-value (? = true) [1][0]] H05-Ans)
        set H05-Ans Ans-list
        set H05-Raw length (remove 0 Ans-list)

        set Ans-list (map [ifelse-value (? = true) [1][0]] H06-Ans)
        set H06-Ans Ans-list
        set H06-Raw length (remove 0 Ans-list)

        set Ans-list (map [ifelse-value (? = true) [1][0]] S05-Ans)
        set S05-Ans Ans-list
        set S05-Raw length (remove 0 Ans-list)

        set Ans-list (map [ifelse-value (? = true) [1][0]] S06-Ans)
        set S06-Ans Ans-list
        set S06-Raw length (remove 0 Ans-list)

        ; these are the codes for the domain raw scores
        set Reading-Raw R04-Raw + R05-Raw + R06-Raw
        set Math-Raw M04-Raw + M05-Raw + M06-Raw
        set History-Raw H05-Raw + H06-Raw
        set Science-Raw S05-Raw + S06-Raw
         ]


end ; end of test routine
```

Appendix B: Analysis of Variance for Data Processing

Due to a concern raised on whether data processing of the Real World data may have affected the natural distribution, an analysis of variance (ANOVA) for the different levels of processing was done. Raw scores were used in this analysis instead of scale scores, which represents a transformation of the raw scores and therefore would alter the distribution according to the transformation. Even though not shown, the results of an ANOVA using the scale scores do not differ. The ANOVA compares distribution at each level of processing to indicate if they are different. Groups 0, 1, and 2 represent the Raw, Complete, and Longitudinal dataset, respectively.

Between-Subjects Factors

		N
GROUP	0	192239
	1	182246
	2	139062

Descriptive Statistics

Dependent Variable: M_RAW

GROUP	Mean	Std. Deviation	N
0	41.18	11.145	192239
1	41.61	10.887	182246
2	42.74	10.318	139062
Total	41.76	10.853	513547

Tests of Between-Subjects Effects

Dependent Variable: M_RAW

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Obs'd Power [a]
Corrected Model	204213[b]	2	102106	870	0.000	0.003	1.000
Intercept	881351255	1	881351255	7507904	0.000	0.936	1.000
GROUP	204213	2	102106	870	0.000	0.003	1.000
Error	60284824	513544	117				
Total	955876786	513547					
Corrected Total	60489037	513546					

Computed using alpha = .05

R Squared = .003 (Adjusted R Squared = .003)

Notice that while the results indicate that data processing does significantly affect the distribution, the effect size is incredibly small. The significant results are due to the high statistical power available from such a large dataset. The mean difference between the Longitudinal and Raw dataset is 1.56 items while the standard deviation is over 10 items. It is reasonable to conclude then that the data processing does not affect the population distribution very much.

Appendix C: Analysis of Variance for Domain Difficulty

An analysis of variance was done to determine whether the different domains were statistically different when it came to difficulty. The coding system 0f 1, 2, 3, and 4 corresponds to Reading, Math, History, and Science respectively. The results of the analysis are below:

Between-Subjects Factors

		N
Domain	1.00	129
	2.00	168
	3.00	105
	4.00	110

Tests of Between-Subjects Effects

Dependent Variable: unstandardized B and theta estimation by section

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Obs'd Power (a)
Corrected Model	47(b)	3	16	55	.000	.246	165.819	1.000
Intercept	306	1	306	1090	.000	.682	1090.479	1.000
Domain	47	3	16	55	.000	.246	165.819	1.000
Error	143	508	.281					
Total	496	512						
Corrected Total	189	511						

a Computed using alpha = .05
b R Squared = .246 (Adjusted R Squared = .242)

Multiple Comparisons

Dependent Variable: unstandardized B and theta estimation by section

	(I) Domain	(J) Domain	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Upper Bound	Lower Bound
Tukey HSD	1	2	-0.68	0.06	0.00	-0.84	-0.52
		3	-0.32	0.07	0.00	-0.50	-0.14
		4	-0.74	0.07	0.00	-0.92	-0.57
	2	1	0.68	0.06	0.00	0.52	0.84
		3	0.36	0.07	0.00	0.19	0.53
		4	-0.06	0.07	0.77	-0.23	0.11
	3	1	0.32	0.07	0.00	0.14	0.50
		2	-0.36	0.07	0.00	-0.53	-0.19
		4	-0.43	0.07	0.00	-0.61	-0.24
	4	1	0.74	0.07	0.00	0.57	0.92
		2	0.06	0.07	0.77	-0.11	0.23
		3	0.43	0.07	0.00	0.24	0.61
LSD	1	2	-0.68	0.06	0.00	-0.80	-0.56
		3	-0.32	0.07	0.00	-0.46	-0.18
		4	-0.74	0.07	0.00	-0.88	-0.61
	2	1	0.68	0.06	0.00	0.56	0.80
		3	0.36	0.07	0.00	0.23	0.49
		4	-0.06	0.07	0.34	-0.19	0.07
	3	1	0.32	0.07	0.00	0.18	0.46
		2	-0.36	0.07	0.00	-0.49	-0.23
		4	-0.43	0.07	0.00	-0.57	-0.28
	4	1	0.74	0.07	0.00	0.61	0.88
		2	0.06	0.07	0.34	-0.07	0.19
		3	0.43	0.07	0.00	0.28	0.57

Based on observed means.

Domain was found to be a significant factor. Post hoc test indicates that Math and Science are not significantly different while Reading and History are significant different to each other and to Math and Science. Based on the mean difficulty, it was concluded that Reading was the easiest domain, then History, with Math and Science being equal and the hardest.

References

Allen, M.J. and Yen, W.M. (1979). **Introduction to Measurement Theory**. Brooks/Cole. Monterey, CA.

Anderson J.C. and Gerbing, D.W. (1988). Structural Equation Modeling in Practice: A Review and Recommended Two Step Approach. **Psychological Bulletin**. 103: 411-423.

Baker, F. (2001). **The Basics of Item Response Theory**. College Park, MD: Univeristy of Maryland.

Birnbaum, A. (1968) Some Latent Trait Models and Their Use in Inferring an Examinee's Ability. In Lord, F.M. and Novick, M.R. (Eds.) **Statistical Theories of Mental Test Scores**. Reading, MA: Addison-Wesley

Bobrow, J. (2004). **Cliffs TestPrep TAKS**. John Wiley & Sons. Hoboken, NJ.

Boslaugh, S. (ed.) (2007). **Encyclopedia of Epidemiology**. New York City, NY: Sage Publishing.

Brooks, S.S. (1922). **Improving Schools by Standardized Tests**. Cambridge, MA: The Riverside Press.

Byth, K. and Cox, D.R. (2005). On the relation between initial value and slope. **Biostatistics**. 6(3) pp. 395-403

Chambers, W.V. (2000). Causation and Corresponding Correlations. **Journal of Mind and Behavior**. 21:437-460.

Chen, G.; Glen, D.R.; Stein, J.L; Meyer-Lindenberg, A.S.; Saad, Z.S.; and Cox, R.W. (2007). Model Validation and Automated Search in FMRI Path Analysis: A Fast Open Source Tool for Structural Equation Modeling. **Human Brain Mapping Conference 2007**. Chicago, IL.

Dewey, J. (1916). **Democracy and Education: An Introduction to the Philosophy of Education**. New York, NY: Macmillan Company

Ellis, J.D. (2003). The Influence of the NationalScience Education Standards on the Science Curriculum. In Hollweg, K.S and Hill, D. (Eds.) **What is the Influence of the National Science Education Standardards? Reviewing the Evidence, a Workshop Summary**. Washington, D.C.: The National Academic Press.

Gonzales, P.; Williams, T.; Jocelyn, L.; Roey, S.; Kastberg, D.; and Brenwald, S. (2008). *Highlights from TIMSS 2007: Mathematics and Science Achievement of U.S. Fourth- and Eighth-Grade Students in an International Context* (NCES 2009-001). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, D.C.

Hashway, R. M. (1998). **Error Free Mental Measurements: Applying Qualitative Item Response Theory to Assessment and Program Validation Including a Developmental Theory of Assessment**. Bethesda, MD: Austin & Winfield.

Ho, R. (2006). **Handbook of Univariate and Multivariate Data Analysis and Interpretation with SPSS**. Boca Raton, FL: Chapman & Hall/CRC

Kane, M. (2006). Content-Related Validity Evidence in Test Development. In Downing, S.M., and Haladyna, T.M (Eds.) **Handbook of Test Development**. Mahwah, NJ: Lawrence Erlbaum

Kelloway, E.K. (1995). Structural Equation Modeling in Perspective. Journal **of Organizational Behavior**. 16: 215-224.

Kline, P. (1993). **Handbook of Psychological Testing**. New York, NY: Routledge.

Kline, T. (2005). **Psychological Testing: A Practical Approach to Design and Evaluation**. Sage Publishing

Lord, F.M and Novick, M.R. (1968). **Statistical Theories of Mental Test Scores**. Reading, MA: Addison-Wesley Publishing Company.

Meyers, J.L., Miller, G.E., and Way, W.D. (2009). Item Position and Item Difficulty Change in an IRT-Based Common Item Equating Design. **Applied Measurement in Education**. 22: 38-60

Micceri, T. (1989). The Unicorn, the Normal Curve, and Other Improbably Creatures. **Psychological Bulletin**. 105 (1) 156-166.

Mueller, R.O. (1997). Structural Equation Modeling: Back to Basics. **Structural Equation Modeling**. 4: 353-369.

Pearson Educational Measurement. (2005). *Pearson Educational Measurement Wins Texas Student Assessment Contract*. Retrieved June 15, 2009 from the World Wide Web: http://www.pearsoned.com/pr_2005/061505.htm

Popham, W.J. (2007). *Instructional Insensitivity of Tests: Accountability's Dire Drawback*. American Educational Research Association: Chicago, IL April 9-13.

Rasch, G. (1980). **Probabilistic Models for Some Intelligence and Attainment Tests**. Reprint of 1960 with Fore and Afterword by Benjamin D. Wright. Chicago, IL: University of Chicago Press.

Rogers, E.M. (1995). **Diffusion of Innovations**. (4th ed.) New York City, NY: Free Press.

Rudner, L.M. (2007). *PARAM-1PL Calibration Software for the 1 Parameter Logistic IRT Model (freeware).* Available: http://edres.org/irt/param

Savage, C.W. and Ehrlich, P. (1992). **Philosophical and Foundational Issues in Measurement Theory**. Hillsdale, NJ: Lawrence Erlbaum Associates.

Smith, G. and Smith, J. (2005). Regression to the Mean in Average Test Scores. **Educational Assessment**. 10(4) pp. 377-399

Stevens, S.S. (1946). On the theory of scales of measurement. **Science**.103 pp. 677-680.

Stroup, W.M., Pham, V.H., and Alexander, C. (2007). *Richardson MathForward Project Second Year Final Report: Math TAKS Results*. Austin, TX: The University of Texas at Austin.

Taris, T. (2000). **A Primer in Longitudinal Data Analysis**. London, England: Sage Publishing.

TEA Student Assessment Division. (2005). *A Study of the Correlation between Course Performance in Algebra I and Alegbra End-of-Course Test Performance.* Retrieved June 15, 2009 from the World Wide Web: http://ritter.tea.state.tx.us/student.assessment/resources/studies/correlation.pdf

TEA Student Assessment Division (2006). 2006 Released TAKS Exam Answer Key. Retrieved June 15, 2009 from the World Wide Web: http://ritter.tea.state.tx.us/student.assessment/resources/release/taks/2006/grxltaksjulykey.pdf

TEA Student Assessment Division. (2006). **Technical Digest 2005-2006**. Retrieved June 15, 2009 from the World Wide Web: http://www.tea.state.tx.us/index3.aspx?id=4391&menu_id3=793

U.S. Department of Education. (2008). *A Nation Accountable: Twenty-five Years After A Nation At Risk*. Washington, D.C.

van der Linden, W.J. and Hambleton, R.K. (Eds.) (1997). **Handbook of Modern Item Response Theory**. New York City, NY: Springer-Verlag New York Inc.

Warton, D.I. (2006). Robustness to Failure of Assumptions of Tests for a Common Slope Amongst Several Allometric Lines – a Simulation Study. **Biometrical Journal**. 49 (2): 286-299.

Wragg, E.C. and Wragg, T. (1997). **Assessment and Learning**. Oxford, UK: Routledge.

Wright, B.D. and Linacre, J.M. (1987). Rasch Model derived from Objectivity. **Rasch Measurement Transaction**. 1(1) 5-6.

Vita

Vinh Huy Pham was born on July 2, 1979 in Saigon, Vietnam and immigrated with his family to the United States on June 30, 1981 as political refugees where he now resides. From an early age, he has shown a keen interest in teaching and working with children but his gift in the academics swayed him into the realm of science rather than education. Accordingly, he earned two Bachelor of Science degrees from Mercer University in Macon, Georgia: one in Biology and one in Chemistry in 2001. Working as a molecular biologist researcher for the next three years, he felt out of place and finally in 2004, realizing his own passion to teach and desire to help children, Vinh finally enrolled in the Science and Mathematics Education program at the University of Texas – Austin. This represented a merging between his science training and his passion for education. He hopes to contribute to the well being of children through his research not only in the United States but across the world. He is engaged to Kendall Reynolds with whom he holds a national title in Country Western dancing and is a professional dance instructor. His other hobbies include gardening and fishing.

Permanent Address: 13614 Oak Pebble, San Antonio, TX 78232

This manuscript was typed by the author.